A DIGITAL LIBRARY APPROACH TO THE RECONSTRUCTION OF ANCIENT SUNKEN SHIPS

A Dissertation

by

CARLOS A. MONROY COBAR

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2010

Major Subject: Computer Science

A DIGITAL LIBRARY APPROACH TO THE RECONSTRUCTION OF ANCIENT SUNKEN SHIPS

A Dissertation

by

CARLOS A. MONROY COBAR

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Richard Furuta
Committee Members,	Filipe Castro
	Frank M. Shipman, III
	Eduardo Urbina
Head of Department,	Valerie Taylor

August 2010

Major Subject: Computer Science

ABSTRACT

A Digital Library Approach to the Reconstruction of Ancient Sunken Ships.

(August 2010)

Carlos A. Monroy Cobar, B.S., Universidad Rafael Landívar, Guatemala;

M.C.S., Texas A&M University

Chair of Advisory Committee: Dr. Richard Furuta

Throughout the ages, countless shipwrecks have left behind a rich historical and technological legacy. In this context, nautical archaeologists study the remains of these boats and ships and the cultures that created and used them. Ship reconstruction can be seen as an *incomplete jigsaw reconstruction problem*. Therefore, I hypothesize that a computational approach based on digital libraries can enhance the reconstruction of a composite object (ship) from fragmented, incomplete, and damaged pieces (timbers and ship remains).

This dissertation describes a framework for enabling the integration of textual and visual information pertaining to wooden vessels from sources in multiple languages. Linking related pieces of information relies on query expansion and improving relevance. This is accomplished with the implementation of an algorithm that derives relationships from terms in a specialized glossary, combining them with properties and concepts expressed in an ontology.

The main archaeological sources used in this dissertation are data generated from a 17th-century Portuguese ship, the *Pepper Wreck*, complemented with information obtained from other documented and studied shipwrecks. Shipbuilding treatises spanning from the late 16th- to the 19th-centuries provide textual sources along with various illustrations. Additional visual materials come from a repository of photographs and drawings documenting numerous underwater excavations and surveys.

The ontology is based on a rich database of archaeological information compiled by Mr. Richard Steffy. The original database was analyzed and transformed into an ontological representation in RDF-OWL. Its creation followed an iterative methodology which included numerous revisions by nautical archaeologists. Although this ontology does not pretend to be a final version, it provides a robust conceptualization.

The proposed approach is evaluated by measuring the usefulness of the glossary and the ontology. Evaluation results show improvements in query expansion across languages based on Blind Relevance Feedback using the glossary as query expansion collection. Similarly, contextualization was also improved by using the ontology for categorizing query results. These results suggest that related external sources can be exploited to better contextualize information in a particular domain. Given the characteristics of the materials in nautical archaeology, the framework proposed in this dissertation can be adapted and extended to other domains.

DEDICATION

To Eveline, Pietro, my mother Martha, and my family

ACKNOWLEDGEMENTS

When I started graduate studies at Texas A&M, my goal was to pursue a master's degree in computer science. Close to the completion of my degree, Dr. Richard Furuta, my academic advisor, asked me if I was interested in joining the Ph.D. program. At that time, there was the possibility of joining the Cervantes Project as a Research Assistant. Looking back after almost seven years, I never dreamed that the journey I embarked on in 2001 would take me into an exciting voyage through the academic and research worlds. His knowledge and vision were instrumental in my academic experience. Thanks Dr. Furuta for sparking in me this interest.

Throughout my years as a doctoral student I was very fortunate to work in numerous digital libraries projects. I am very thankful for what I learned from Dr. Enrique Mallen (On-line Picasso Project), Dr. Gary Stringer (Digital Donne), Dr. Eduardo Urbina (Cervantes Project), Dr. Filipe Castro (Nautical Archaeology), and Dr. Frank Shipman. Special thanks go to the last three individuals and Dr. Furuta for being on my committee. Collaborating with scholars in other disciplines turned out to be an extremely stimulating environment. Thanks go to Dr. Filipe Castro, whose knowledge about nautical archaeology and enthusiasm for the use of new technologies were essential in completing my dissertation.

My doctoral studies would have been impossible without the financial support of numerous institutions. A Fulbright scholarship made possible my master's studies. My doctoral studies were primarily funded through grants from The National Science

Foundation, The National Endowment for the Humanities, various Computer Science Department scholarships, and a scholarship from the Glasscock Center for the Humanities, thanks Dr. Jim Rosenheim. Dr. Gary Stringer, through the Donne Project, generously supported a couple of conferences I attended and a scholarship.

When it was time to choose a dissertation topic, a casual conversation with my friend, Wendy van Duivenvoorde, a doctoral candidate in nautical archaeology at that time, led to a more formal meeting with Drs. Castro and Furuta. This in turn evolved into an NSF-funded project: The Nautical Archaeology Digital Library. The availability of the materials is the work of numerous people, including Heather Hatch and Kevin Gnadinger, who worked on the edition of *glosShip* and digitization of shipbuilding treatises.

The Cushing Memorial Library at Texas A&M acquired and made available priceless original shipbuilding treatises. The National Library of Portugal granted permissions to use copies and transcriptions of some of the most important Portuguese shipbuilding treatises. Jean-Michel Urvoy worked on *glosShip* and shared his knowledge in naval architecture. Mauro Bondioli's conversations regarding the glossary were very fruitful; he raised numerous questions about my approach. Thanks Mauro.

In 2008 I participated in the Nebraska Digital Workshop. Presenting my work and sharing my research with a group of scholars in digital humanities was a great experience. Special thanks go to Dr. Gregory Crane of the Perseus Project. He not only critiqued my project but also gave me numerous ideas, and expanded my vision on the potential of my research.

This dissertation, especially *ontoShipDS*, would have been impossible without the visionary work of Mr. Richard Steffy. His compilation of information and data pertaining to decades of research at CMAC, and the creation of a database for storing them is priceless. Mr. Steffy's inquisitive spirit can almost be touched when reading his book on nautical archaeology.

Special thanks go to the current Department of Computer Science Head, Dr. Valerie Taylor. Although she has a busy schedule, her doors were open anytime I requested an appointment, in particular when I discussed with her an academic collaboration with Universidad Rafael Landivar—my Alma Mater in Guatemala. The fruit of that conversation was Dr. Bjarne Stroustrup graciously giving a videoconference talk to two universities in Guatemala. Thanks Dr. Stroustrup. I found my conversations with other computer science faculty very stimulating throughout my years as a student, in particular those with Drs. Ioerger, Gutierrez, Kerne, and Caverlee.

I spent numerous hours with other graduate students at the Center for the Study of Digital Libraries. Conversations with them or a quick trip to get coffee on the third floor made me realize how fortunate I was to be surrounded by great people. Our weekly meetings were an opportunity for ideas and discussions that enriched my academic life. I had to commute from Houston during my last year at A&M. Commuting with Zach Toups made the trip not only more enjoyable, but talking to him and learning from his research in Computer Human Interaction was enriching.

In the spring of 2009, I joined the Center for Technology in Teaching and Learning at Rice University in Houston. Working in a multidisciplinary team (educators, scientists, digital artists, and psychologists) expanded my horizons by participating throughout the life-cycle of games for education (conceptualization, development, deployment, and evaluation). I am grateful to Dr. Leslie Miller for opening my eyes to the potential of games in education and digital humanities.

St. Mary's Catholic Center was (and is) a great place for *recharging* the batteries. I made numerous friendships throughout my years at Texas A&M. Thanks Fr. Mike Sis for your friendship. Your example and simplicity is something I will treasure all my life. Fr. David Konderla shared his wisdom, serenity, and sense of humor.

Numerous friends and relatives scattered in different countries made all these years away from my beloved Guatemala, a worthy experience, in particular my dear friend, Juan Roberto Brenes, who died in 2009. I am thankful to Centro Universitario Ciudad Vieja, for the priceless education I received and the friendships I made during my undergraduate years. Finally, thanks go to my mother, Martha, for her example and encouragement, to my wife, Eveline, for her companionship and love, and to my son, Pietro, for brightening my life in a very special way. He was born when I was in the last stages of my dissertation (which changed my life dramatically). Although sleepless nights were common, contemplating a simple smile from him made me realize that his presence has a deep meaning in my life.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

Page

Page

LIST OF TABLES

1. INTRODUCTION

The research involved in this dissertation lies at the intersection of two disciplines: nautical archaeology and digital libraries. Nautical archaeology as science is the study of the remains of boats and ships and the cultures that created and used them.

In recent years, archaeologists have been gradually incorporating computers, specialized software, and ad-hoc databases to support their work. However, numerous challenges still remain. In nautical archaeology two of the main problems arise from the characteristics of the objects under study and the complexity of the tasks undertaken.

This dissertation focuses on ship reconstruction, in particular in the creation of a framework for integrating and augmenting supporting materials. Rebuilding a vessel can be seen as an *incomplete jigsaw reconstruction problem*, where fragments and pieces have to be identified, tested, and placed together. The difficulties associated to this endeavor are summarized by Dick Steffy when describing the information generated by archaeological excavations:

> Those disks contain a wealth of information—excavation catalogs, texts of articles and books, a hull data for dozens of ships, correspondence, and so on. It has taken many years to accumulate all that information, yet it is unrecognizable and inaccessible unless one knows the proper codes. Shipwrecks are like that. Piles of rotted timbers and broken artifacts constitute a wealth of information, yet much of that knowledge will remain unrecognized unless one develops a proper method of access to it. (p. 189).

My main hypothesis can be stated as follows: a computational approach based on

---

This dissertation follows the style of *ACM Transactions on Information Systems.*

digital libraries can enhance the reconstruction of a composite object (ship) from fragmented, incomplete, and damaged pieces (timbers and ship remains).

Supporting materials employed in ship reconstruction can be divided into textual and visual. Textual sources pertain to written technical documents in multiple languages that describe ship components and assembly instructions. Visual or graphic sources on the other hand, refer to technical illustrations, drawings, and photographs. Thus, the problem can be summarized as the creation of a method capable of indexing, retrieving, and associating relevant pieces of information.

Digital libraries and information retrieval are two of the disciplines related to the problem tackled in this dissertation. A full-text index for example, facilitates the storage and retrieval of information in the collection. Numerous information retrieval techniques help to improve finding relevant information in a multilingual environment. One way to accomplish this is by incorporating relevance feedback in the retrieval process. This is an important feature because it narrows down retrieved information by ranking higher relevant results, and discarding less relevant ones (indicated by lower rankings).

There exist numerous relevance feedback methods. Some use an external collection (frequently a general purpose one) for query expansion. It has been demonstrated that this approach improves relevance [Xu and Croft 1996]. Because nautical archaeology is a specialized domain, a glossary of nautical terms is used as external collection for evaluating query expansion (Section 8.3).

Although useful for query expansion, a glossary has limitations in expressing relationships among terms. A solution to this shortcoming is the incorporation of a specialized ontology [Soo et al. 2002]. Ontologies have been extensively used for describing concepts, properties, and relationships in numerous domains (see the next section). Thus the ontology further improves contextualization by exploiting additional information not included in the glossary. This is discussed in Section 8.3.

The dissertation is organized as follows; section three offers a glimpse into Nautical Archaeology, the process of ship reconstruction, and an introduction to the *Pepper Wreck*. Section four introduces textual materials—shipbuilding treatises—and elaborates on how they have been structured. It also describes *glosShip*—a multilingual glossary of nautical terms, along with information retrieval techniques used for improving relevance, in particular the use of Blind Relevance Feedback. Because textual materials are enriched with graphic and visual resources, section five discusses illustrations, photographs, and drawings and their role in ship reconstruction.

Section six addresses the creation of *ontoShipDS*—an ontology for describing wooden ships. *ontoShipDS* is based on the pioneering work of Mr. Dick Steffy. In order to explain how the different components of the proposed framework are integrated, section seven covers examples of uses of the ontology with both textual and visual resources. Tests and results in the evaluation of my approach are presented in section eight. Finally, section nine concludes the dissertation with a discussion about the lessons learned in this work, and some ideas for future work.

2. RELATED WORK

Although the principles and methods used in archaeological excavations have not changed much over the last decades [Renfrew and Bahn 2000; Ashmore and Sharer 2003; Thomas 1998] advances in computing, recording, and instrumentation technology have impacted both the way off-the-field work is carried out and how the gathered data is stored, processed, and presented. Mapping tools and customizable databases, for instance, have transformed the day to day work of archaeologists. Software such as HPASS [Green and Souter 2002], SITE SURVEYOR [Holt 2008], PhotoModeler [Green et al. 2002], and Rhinoceros [Rhinoceros 2009] illustrate examples ranging from GPS, triangulation, CAD, to 3D modeling used in nautical archaeology.

The *Petra Great Temple excavations* [Joukowsky 1993; Egan et al. 2000] is a joint archaeological excavation in Jordan that illustrates the development of new technologies based on the archaeologists needs. The *Brown University SHAPE Laboratory* [SHAPE 2008; Hadingham 2000; Acevedo et al. 2000; Acevedo et al. 2001] has developed various techniques and tools applied to Archaeology; such as 3D-free-form models for geometric recovery, virtual environments, topography, and linguistics elements. The *Theban Mapping Project* [Theban Mapping Project 2008; Reeves 1992; Reeves and Wilkinson 1996] provides a comprehensive archaeological detailed map and database of every archaeological, geological, and ethnographic feature in Thebes.

Within the scope of digital humanities, numerous scholars are embracing new technologies for the preservation, dissemination, and exploration of a wide range of collections. The *Perseus Project* [Crane 1996; Crane 1988] is an example of a digital library in the context of cultural and historical heritage material. Focused originally on ancient Greek culture, and currently including Roman and Renaissance collections, it provides a variety of visualization tools and navigation options for a collection of texts and images. The *Digital Atheneum* [Brown and Seales 2000; Brown and Seales 2001] has developed new techniques for restoring, searching, and editing humanities collections. The *Digital Imprint* [The Digital Imprint 2008] proposes design standards for the electronic publication of archaeological site reports.

With the idea of reconstructing the history of the text, the *Canterbury Tales Project* [The Canterbury Tales Project 2008] has collated several of the manuscripts written by Chaucer. The *Rossetti Archive* [The Rossetti Archive 2006] is a hypertextual-based collection of the pictorial and textual works of Dante Gabriel Rossetti. *Digital Donne* [Monroy et al. 2007] provides a collection of poems of the British poet John Donne, including a comprehensive collation of various copies of both printed and handwritten copies.

Reconstructing an ancient and no-longer-existing object or environment is a process subject to ambiguity. Archeologists' decisions are often based on limited and damaged physical evidence, similar existing objects, or textual references [Kensek et al. 2004, p. 176]. One way to address the information used in the reconstruction of a physical object (temple, artifact, or ship), is to ensure that supporting evidence can be linked to the virtual object as a way to document the decisions made [Snyder 2004]. In order to offer a meaningful interpretation of a reconstruction, [Snyder 2008] also suggests the incorporation of expert commentary, primary sources, and graphic references.

Similarly, ship reconstruction can be based on inferences from partially recovered archaeological evidence and other written sources [Castro 2005c; Van Duivenvoorde 2009]. This is the case in the excavation of the Portuguese ship *Santo Antonio de Tanná*, whose length and breadth had to be inferred from surviving portions of the hull and parts of the frames, along with historical sources [Fraga 2007, p. 178]. These interpretations can be complemented with construction sequences described in textual sources such as shipbuilding treatises. A methodology for representing ambiguity in a non-intrusive way is proposed by [Kensek et al. 2004]. With special emphasis on ambiguity in 2D reconstructions, the authors link textual and visual information for documenting decisions made by archaeologists.

In this context [Forte 1997] states that one of the goals of virtual archaeological reconstructions is "to enhance and direct cognitive perceptions of antiquity." However, [Kensek et al. 2004] argue about the risk associated with these reconstructions when supporting materials and evidence are not properly referenced:

> The risk is that virtual archaeological reconstructions can have lives of their own; they are seductive; they can seem viable. This situation, especially when there is no explicit linkage to an evidence corpus of the type usually provided in text-based or static 2D reconstructions in the form of footnotes, or when the virtual product is not properly constrained because of the lack of evidence, or when a virtual reconstruction is created without rigorous reference to the extant evidence sources. (p. 175).

In this context, [Blomerus and Lesk 2008; Lesk 2004] propose a method for linking a database to a CAD model in the reconstruction of a Greek building. Their contribution is a 4D CAD model, in which time as the fourth dimension enables visualization of the building at different time periods.

In nautical archaeology, according to [Steffy 1994], ship reconstruction can take three forms: graphic, three-dimensional, and physical. Graphic reconstructions are primarily two-dimensional. However, regardless of the type of reconstruction, the final product includes reports composed of texts, photographs, and drawings. A physical reconstruction, on the other hand, entails rebuilding the real full size vessel from the recovered and preserved timbers. Although, this is the pinnacle of ship reconstruction, it tends to be extremely expensive, complex, and time consuming.

A common characteristic in the previous examples is the ability to connect parts of a rebuilt physical object with relevant sources or materials. Following textual descriptions in various shipbuilding treatises, [Hazlett 2007] created a 3D model of a 16th–century Portuguese ship. The author eloquently captures the essence of his approach by calling it a "textual excavation." Although the reconstruction was primarily based on two written sources, it was complemented with additional texts, graphic materials, and archaeological evidence. The author summarizes his approach working with a multilingual collection as follows:

> Part of the reconstruction work presented below was my own English-language translation of one of the documents in the 1590 *Livro Nautico*… My translation is based upon Mendonça's 1898 transcription of this document into modern Portuguese… I translated Mendonça's Portuguese text into English, and checked it against images of the original handwritten document. (p. 48).

Unlike Hazlett's work, [Wells 2008] created a 3D reconstruction of a Portuguese vessel based on interpreted archaeological evidence in the form of drawings. The author's iterative approach shows how new interpretations can affect the digital model. This is summarized as follows:

> The archaeological remains of the ship have been excavated and interpreted, resulting in data (such as measurements of the hull remains) and theory (such as proposed reconstructed hull dimensions). The data and theory inform the creation of the digital model, which can then be used to make further interpretations. These new interpretations can bring about changes in the data, theory, and digital model indefinitely. (p. 2).

In archaeology, the importance of associating the objects under study is highlighted by [Doerr and Crofts 1999]:

> Archaeologists and paleontologists habitually deal with fragmented objects, which are then combined, with luck, into a single whole—a process that is highly unusual in other domains. Multiple fragments need to be identified and tracked during the entire process. (p. 3).

In technical and specialized domains such as nautical archaeology, linking materials can be even more difficult given the media and language of the sources. Examples of this include photographs and technical illustrations partially showing sections of a ship, as well as shipbuilding treatises written in different languages, where the terms and concepts can change over time (Mauro Bondioli, personal communication May 2007).

Information retrieval is a discipline that offers methodologies and algorithms for finding relevant objects in a collection. Query expansion for instance, modifies an original query using retrieved documents or additional resources to improve overall retrieval. [Cutler et al. 1999] show retrieval improvements using statistical-based methods for query expansion. A methodology based on anchor text analysis shows that anchor text can be used to improve precision [Eiron and McCuley 2003]. [Vechtomova and Wang 2006] evaluated various distance functions for selecting query expansion terms, showing that distance-weighted mutual information outperforms mutual information alone.

In a cross language environment, [Adriani and van Rijsbergen 1999] demonstrate the use of term-similarity-based techniques to improve dictionary-based retrieval in a multilingual collection. Based on experimental results, [Gazendam et al. 2006] describe an approach for ranking metadata based on relationships derived from words in a domain-specific thesaurus in Dutch. [Vechtomova and Karamuftuoglub 2007] show that query expansion based on snippets outperformed whole document-based query expansion.

Comparing various query expansion methods, [Xu and Croft 1996] demonstrated that local contexts analysis based on global analysis techniques produce better results than simple local feedback. On the other hand, [Pérez-Agüera and Zaragoza 2008] combine query expansion and semantic disambiguation for improving ad-hoc retrieval. [Machado et al. 2009] increased performance in a retrieval system using a method base on query expansion and multinomial language modeling. Experimental results based on query reweighting and query expansion showed significant improvements in an interactive information retrieval environment [Harman 1992].

Because of the advantages in representing knowledge, ontologies have been adopted in a wide range of applications in diverse domains. In Artificial Intelligence, the definition of ontology can be summarized as a formal specification of a conceptualization [Gruber 1993; Studer et al. 1998].

Within information retrieval, ontologies can be used to improve retrieval of relevant information. In the legal domain, [Schweighofer and Geist 2007] show improvements employing user behavior along with a legal ontology for improving information retrieval in a collection of legal documents. [Navigli and Velardi 2003] state that deriving semantic information from an ontology outperforms query expansion based solely on synonyms and hypernyms. Exploiting domain and a geographical ontology, [Fu et al. 2005] show improved search results.

An application of ontologies to engineering design is discussed by [Tudorache 2006]. The application of ontologies in Tudorache's work has some similarities with ships. The author applies ontologies in the design of modern machines. Machines are made out of parts, and parts are interconnected. This has some similarities with ships, where sections of the ship can be recursively decomposed in smaller parts.

In Bioinformatics, [Kawazoe et al. 2008] describe an event ontology for detecting infectious disease-related events reported in textual sources in a multilingual environment. Combining a domain ontology and a thesaurus, [Soo et al. 2002] implement two similarity algorithms for improving image retrieval. Based on an ontology that describes characteristics of genes and their products, amiGo [Carbon et al. 2008; Day-Richter et al. 2007] offers a suite of software tools for making available a rich collection of information about genes. [Ciula et al. 2008] propose an ontology for modeling information in historical documents.

3. NAUTICAL ARCHAEOLOGY AND SHIP RECONSTRUCTION

3.1 Brief Definition of Nautical Archaeology

Nautical archaeology as science is the study of the remains of boats and ships and the cultures that created and used them. Underwater excavations include the recovery (when possible) of artifacts and ship components for further analysis. Due to the effects of the environment and time, archaeological evidence has to be carefully treated in order to guarantee long-term preservation. Furthermore, archaeologists destroy archaeological sites as they dig and recover artifacts.

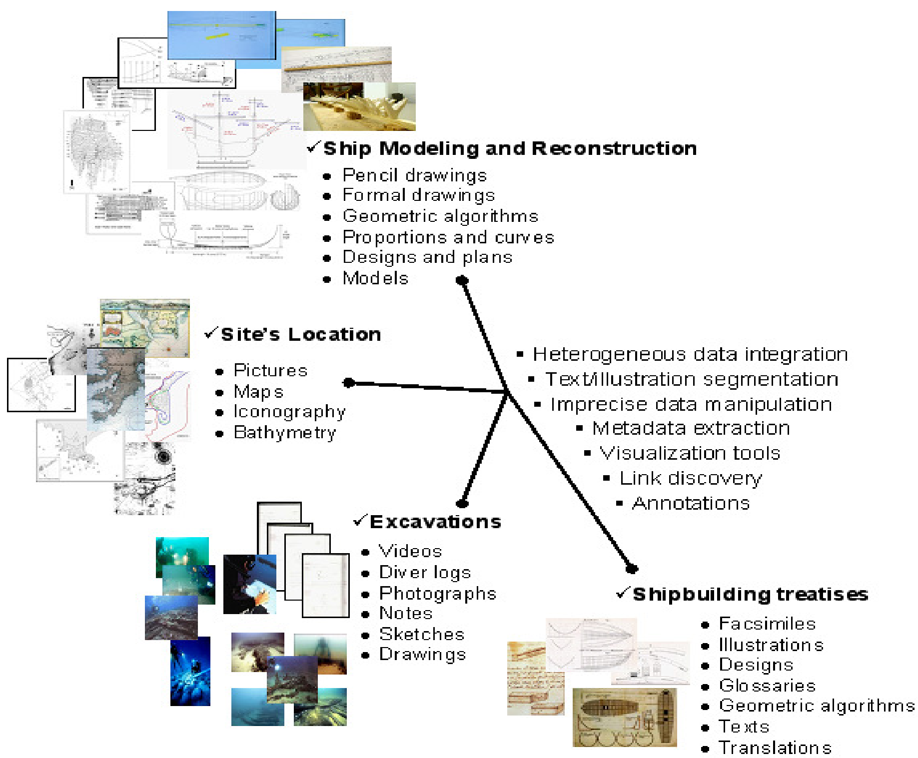


Figure 3.1. Activities, objects, and supporting materials in ship reconstruction.

Archaeologists and scholars studying and analyzing artifacts and ship components produce large amounts of information in a variety of media, such as paper notes, sketches and drawings, pictures in negatives, paper prints, slides or digital images, and videos. In addition, cataloging and documenting this process generates text documents, spreadsheets, databases, and specific file formats resulting from diverse software utilized. Figure 3.1 depicts the major activities in ship reconstruction along with various sources of information.

3.2 Ship Reconstruction

Ships are time capsules and their remains often include hull timbers and rigging elements, often broken and partly destroyed, that can yield precious information about their design and construction. Recording the physical characteristics of these timbers is important for the subsequent study of the ship's remains.

Understanding the archaeology of ships is an important task because these were among the most complex artifacts built by men during many centuries. Understanding the process of conceptualization, design and construction of early ships is an important subfield of the history of technology, and when we consider the ships of the post-medieval period, nautical archaeology becomes part of the history of science.

Reconstructing composite objects—such as ships—from incomplete or damaged sources, involves the combination of algorithmic techniques and visualization tools. This requires locating dynamic information as timbers and objects are constantly recovered and documented. However, this information is not limited to ship parts from the same vessel. Techniques, wood properties, and construction sequences from other construction traditions, help to understand any ship under study for which incomplete archaeological evidence was recovered or no written records are known. In addition, information encoded in shipbuilding treatises (textual and graphic), regardless of their age and provenance, is also valuable to complement the understanding of the archaeological evidence.

Another important aspect in ship reconstruction is the proper classification of types of watercrafts. Ships have traditionally been organized in obvious ways, such as tonnage and rigging type. However, for reconstruction purposes it is important to study and understand other characteristics, mainly related with the way in which they were conceived, designed, and built. For instance, a basic division between shell-based and skeleton-based construction is proposed by [Basch 1972]. In contrast, an approach based on particular traits in the ship's structures (called architectural signatures) is advanced by [Rieth 1984; Rieth 1998].

3.3 The *Pepper Wreck*

Located in São Julião da Barra (Portugal), at the mouth of the Tagus river, the *Pepper Wreck* [Castro 1998; Simonetta 1998; Castro 2003; Castro 2005a; Castro 2005b] site is only a few miles away from Lisbon. Its excavation yielded a large collection of artifacts dated from the late 16th and early 17th centuries, and lead to the identification of this shipwreck as the nau *Nossa Senhora dos Mártires,* wrecked on September 15 1606 on its way back from India. The study of its hull remains—which include a portion of the keel, eleven frames, and some of the planking—yielded interesting results and a first glance at these largely unknown ships: the Portuguese *naus da India*.

The excavation and study of this vessel illustrates how, even with a small number of timbers coupled with written sources, it is possible to conceptualize the entire ship [Castro 2005c].

> It was solidly built of stone pine planks nailed to frames of cork oak with long iron spikes. Using some of its critical dimensions, and applying sets of proportions described in three Portuguese treatises on shipbuilding from the late 16th and early 17th centuries, we tentatively reconstructed the hull. (p. 151).

The following photographs belong to the *Pepper Wreck*, and are intended to give an idea of an underwater archaeological excavation, as well as the recovered ship components. Figure 3.2 shows a damaged and fragmented section of a keel (at the conservation site). Figure 3.3 depicts a diver working on the timbers. Some timbers from the hull of the ship at the excavation site are shown in Figure 3.4. Details of various nails recovered from an excavation are depicted in Figure 3.5.

Figure 3.2. A damaged segment of a keel. Photograph courtesy of CMAC.

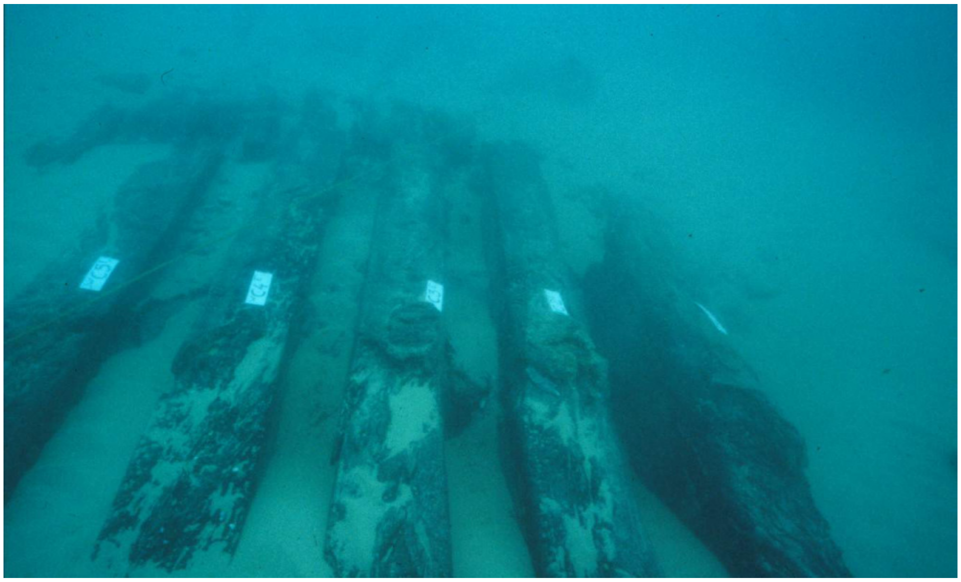


Figure 3.3. Documenting timbers of the hull of the *Pepper Wreck*.
Photograph courtesy of CMAC.

Figure 3.4. A diver working on the hull of the *Pepper Wreck*. Photograph courtesy of CMAC.

Figure 3.5. Nails recovered from the *Pepper Wreck*. Photograph courtesy of CMAC.

4. TEXTUAL ABSTRACTIONS OF SHIPS

4.1 Introduction

Abstractions (understood as representations) of real physical objects can be conveyed in various forms. Photographs, drawings, and illustrations belong to visual abstractions (discussed in Section 5). Written materials, in contrast, correspond to textual abstractions. In this section I discuss a type of texts—known as shipbuilding treatises—whose contents depict descriptions of ship components and frequently provide assembling instructions.

I begin discussing the contents of the treatises and how ship components are described. The second part details the approach I propose for structuring the contents, which in turn enables segmentation, transformations, and aggregation. Linguistic diversity (treatises written in various languages) is one of the characteristics of the materials in this collection. Therefore, I describe the architecture of a multilingual glossary of nautical terms and concepts.

Given the variety of genres, domains, and goals in the study of texts, there are multiple ways in which they can be approached and studied. The abundant literature, conferences, and interest groups in fields such as textual studies, text analysis, and linguistics to name a few, offer a glimpse into the vast universe of texts. Analytical bibliography, for example, studies texts as physical and tangible objects [Greg 1933; Kraft 1990]. The study of the structure and meaning of the language (grammar and semantics) pertains to the domain of linguistics, which in turn can be divided into other categories such as: phonetics, phonology, morphology, syntax, and semantics; to name a few.

With the idea of contrasting two different views on what texts really are, Allen Renear and Jerome McGann discussed the text as a hierarchy of objects and as a performance respectively [Hockey 1999]. From the humanities computing perspective [Rockwell 2003] summarizes this comparative view as follows:

> Renear put forward, for the sake of the confrontation, the OHCO (ordered hierarchy of content objects) perspective while McGann practiced a view of text as performance. In the context of a humanities computing conference this confrontation was designed to highlight the relationship between theories of text and ways of representing texts digitally. Renear's Platonic view of the text as a real abstract object fits nicely with the dominant practice for the digital representation of texts, as represented by the guidelines of the TEI. McGann instead gave us an example of a reading that was both a performance itself and pointed to the combinatorial possibilities within and around the text. McGann's challenge to Renear was to show how a playful reading of a text was both a new text and that this potential could not be captured easily by an OCHO. The confrontation succinctly opened again the question of the relationship between how we represent texts, how we use them, and our theories of textuality. (pp. 209-210).

In this section I examine the textual abstraction of a particular type of composite object: the ship. Textual abstraction in this context means: how the parts of a vessel along their relationships are textually expressed. This representation is important not only because they comprise ship parts, their properties, and how they are related; but also because these specialized ancient technical narratives offer essential construction sequences, hence their relevance in nautical archaeology.

Without written sources, ship reconstruction is a more difficult undertaking. Ship reconstruction of vessels belonging to periods for which shipbuilding treatises were nonexistent or access to these written sources is not possible, relies more on archaeological evidence, and to certain extent on other non-technical historical narratives.

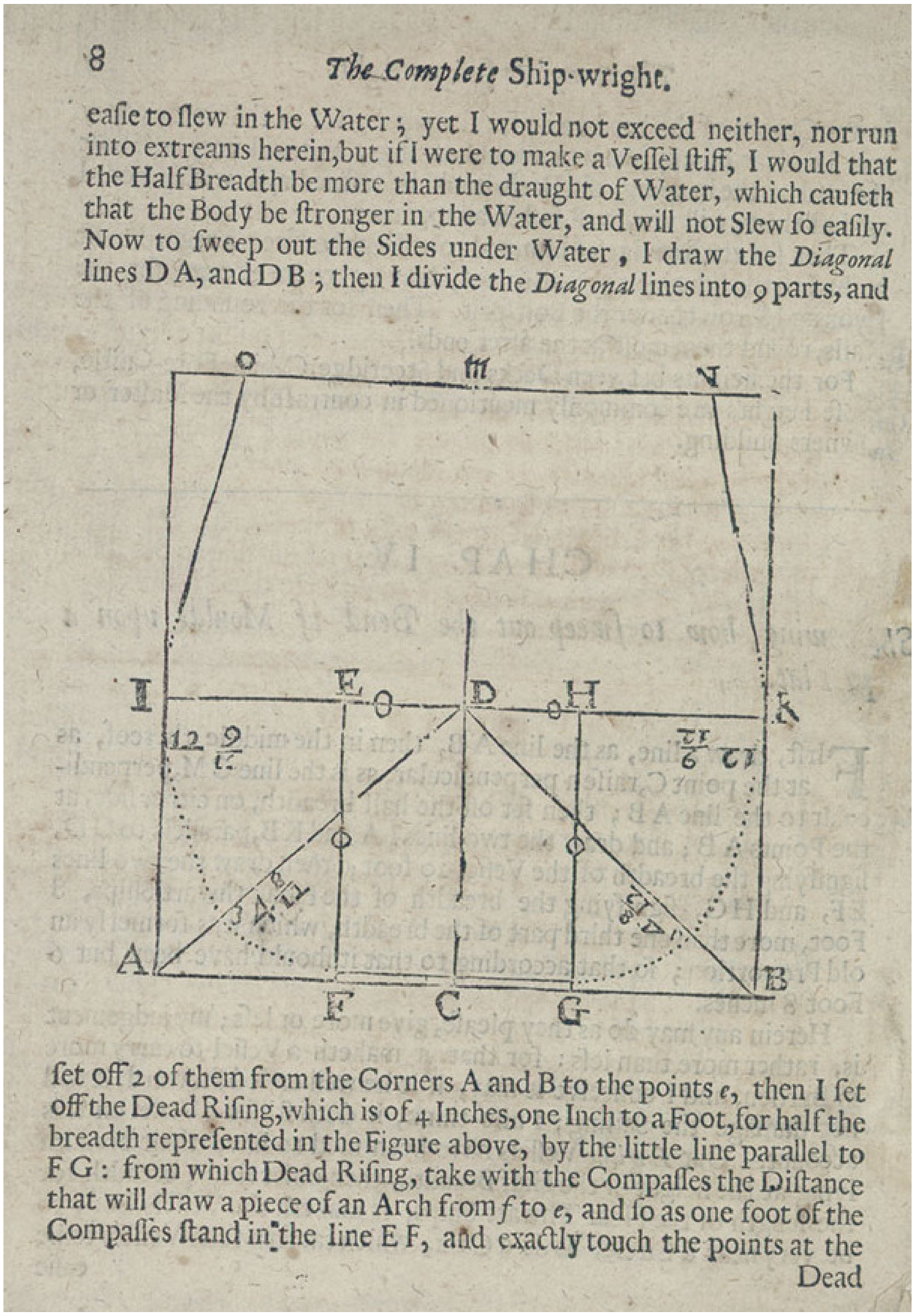
In the context of this dissertation, textual abstractions (represented by shipbuilding treatises) play a dual role in relation to the representation of composite objects. On the one hand, these texts complement visual abstractions of ships. A narrative accompanying an illustration offers a more detailed description of what is graphically shown. Conversely, an illustration depicting a section of a ship offers a visual contextualization of the written text.

4.2 Shipbuilding Treatises

Shipbuilding treatises are technical texts, both printed and manuscript. Although their contents vary, they describe (among other things), the properties and types of wood, methods, measurements, and construction sequences [Hazlett 2006, pp. 69-71]. A digital library infrastructure for storing and presenting shipbuilding treatises is discussed in [Monroy et al. 2007b].

A complete description of the most significant shipbuilding treatises is beyond the scope of this dissertation because each individual work has its own unique characteristics. However, speaking of structuring these written sources would be incomplete without at least an overview of their contents. The analogy I suggest for grasping their contents is to visualize them as technical manuals describing the parts of a machine (ship) and its construction, similar to manuals accompanying modern self-assembling furniture.

Among printed books, Figure 4.1 shows a page from Edmund Bushnell's 1664 *The Complete Shipwright*. Written in English, this treatise includes geometric proportions, arithmetic, and measurements. Because my approach considers the integration of textual and visual abstractions, this figure is a good example depicting both text along with a related illustration. Figure 4.2 is also from Bushnell's book. The colored lines (added by me) are not part of the original document; they were added to highlight part of the contents of a shipbuilding treatise. Red lines correspond to ship components. Green lines underline geometric lines and arcs. Blue lines indicate measurements. Figure 4.3 depicts a partial image of a page from Cornelis van IJk's *De Nederlandsche Scheeps-bouw-konst Open Gestel*, a Dutch treatise published in 1697. Among the information included in this page are a table of proportions for ships of various lengths, and various measurements within the text in the paragraphs. A page (containing text and illustration) from Bartolomeo Crescenzio's *Nautica Mediterranea*, written in Italian and published in 1601 is depicted in Figure 4.4. The text in this example corresponds to geometric proportions of some parts of a ship. What can be observed from this example is the fact that the textual description corresponds to the illustration.

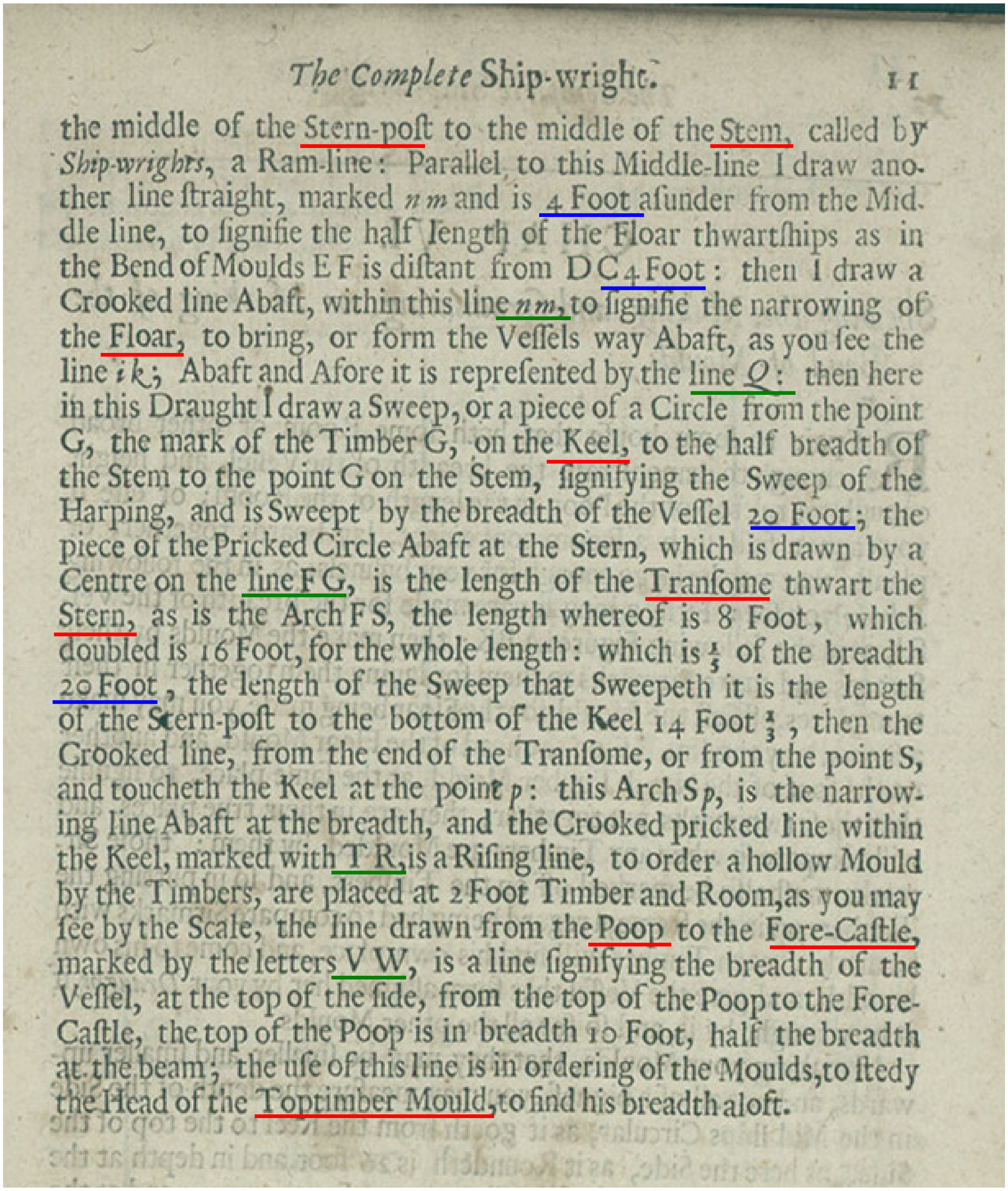


8 *The Complete* Ship-wright.

eaſie to ſlew in the Water; yet I would not exceed neither, nor run into extreams herein, but if I were to make a Veſſel ſtiff, I would that the Half Breadth be more than the draught of Water, which cauſeth that the Body be ſtronger in the Water, and will not Slew ſo eaſily. Now to ſweep out the Sides under Water, I draw the *Diagonal* lines D A, and D B; then I divide the *Diagonal* lines into 9 parts, and

ſet off 2 of them from the Corners A and B to the points *e*, then I ſet off the Dead Riſing, which is of 4 Inches, one Inch to a Foot, for half the breadth repreſented in the Figure above, by the little line parallel to F G: from which Dead Riſing, take with the Compaſſes the Diſtance that will draw a piece of an Arch from *f* to *e*, and ſo as one foot of the Compaſſes ſtand in the line E F, and exactly touch the points at the Dead

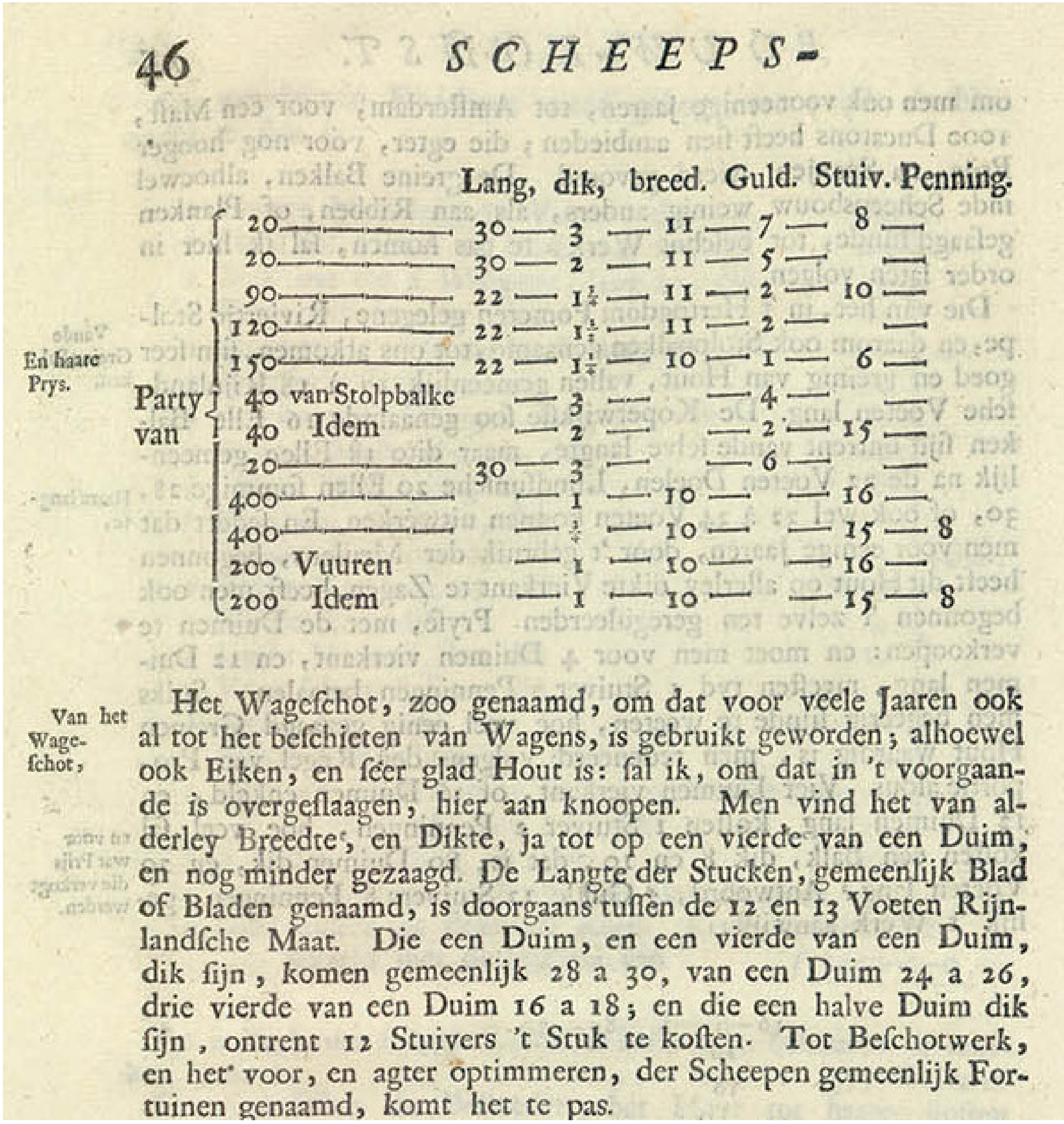
Figure 4.1. Depiction of a page from 1664 Edmund Bushnell's *The Complete Shipwright* (in English).



The Complete Ship-wright. 11

the middle of the Stern-poſt to the middle of the Stem, called by *Ship-wrights*, a Ram-line: Parallel to this Middle-line I draw another line ſtraight, marked *n m* and is 4 Foot aſunder from the Middle line, to ſignifie the half length of the Floar thwartſhips as in the Bend of Moulds E F is diſtant from D C 4 Foot: then I draw a Crooked line Abaft, within this line *n m*, to ſignifie the narrowing of the Floar, to bring, or form the Veſſels way Abaft, as you ſee the line *i k*; Abaft and Afore it is repreſented by the line *Q*: then here in this Draught I draw a Sweep, or a piece of a Circle from the point G, the mark of the Timber G, on the Keel, to the half breadth of the Stem to the point G on the Stem, ſignifying the Sweep of the Harping, and is Sweept by the breadth of the Veſſel 20 Foot; the piece of the Pricked Circle Abaft at the Stern, which is drawn by a Centre on the line F G, is the length of the Tranſome thwart the Stern, as is the Arch F S, the length whereof is 8 Foot, which doubled is 16 Foot, for the whole length: which is $\frac{1}{5}$ of the breadth 20 Foot, the length of the Sweep that Sweepeth it is the length of the Stern-poſt to the bottom of the Keel 14 Foot $\frac{2}{3}$, then the Crooked line, from the end of the Tranſome, or from the point S, and toucheth the Keel at the point *p*: this Arch S *p*, is the narrowing line Abaft at the breadth, and the Crooked pricked line within the Keel, marked with T R, is a Riſing line, to order a hollow Mould by the Timbers, are placed at 2 Foot Timber and Room, as you may ſee by the Scale, the line drawn from the Poop to the Fore-Caſtle, marked by the letters V W, is a line ſignifying the breadth of the Veſſel, at the top of the ſide, from the top of the Poop to the Fore-Caſtle, the top of the Poop is in breadth 10 Foot, half the breadth at the beam; the uſe of this line is in ordering of the Moulds, to ſtedy the Head of the Toptimber Mould, to find his breadth aloft.

Figure 4.2. A page from Bushnell's treatise (colored lines are not part of the original book).



En haare Prys.

	Lang,	dik,	breed.	Guld.	Stuiv.	Penning.
20	30	3	11	7	8	
20	30	2	11	5		
90	22	1¼	11	2	10	
120	22	1½	11	2		
150	22	1¼	10	1	6	
Party van 40 van Stolpbalke		3		4		
40 Idem		2		2	15	
20	30	3		6		
400		1	10		16	
400		¾	10		15	8
200 Vuuren		1	10		16	
200 Idem		1	10		15	8

Van het Wage-ſchot,

Het Wageſchot, zoo genaamd, om dat voor veele Jaaren ook al tot het beſchieten van Wagens, is gebruikt geworden; alhoewel ook Eiken, en ſeer glad Hout is: ſal ik, om dat in 't voorgaande is overgeſlaagen, hier aan knoopen. Men vind het van allerley Breedte, en Dikte, ja tot op een vierde van een Duim, en nog minder gezaagd. De Langte der Stucken, gemeenlijk Blad of Bladen genaamd, is doorgaans tuſſen de 12 en 13 Voeten Rijnlandſche Maat. Die een Duim, en een vierde van een Duim, dik ſijn, komen gemeenlijk 28 a 30, van een Duim 24 a 26, drie vierde van een Duim 16 a 18; en die een halve Duim dik ſijn, ontrent 12 Stuivers 't Stuk te koſten. Tot Beſchotwerk, en het voor, en agter optimmeren, der Scheepen gemeenlijk Fortuinen genaamd, komt het te pas.

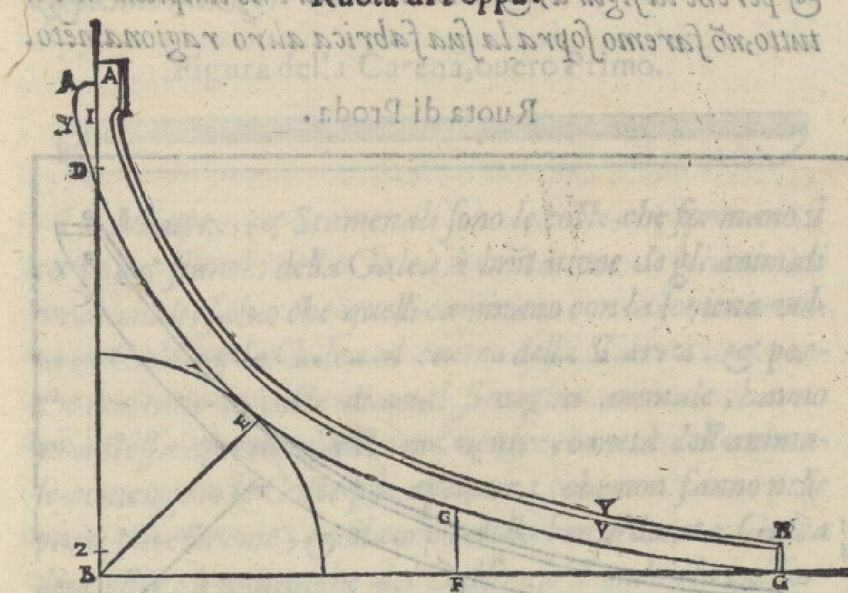
Figure 4.3. Partial depiction of a page from 1697 Cornelis van IJk's *De Nederlandsche Scheeps-bouw-konst Open Gestel* (in Dutch).

Parimente nel piede B.C. annouerando dal B. verſo il C. ſegna palmi vndici notati con la F. per il Calcagnuolo: detto Calcagnuolo è alto palmi 2.

Diuidendo vltimamente l'angolo A.B.C. per mezo à palmi 7. ſegna con l'E. che ſi chiama lo ſgarramento.

Il che fatto, tirerai con la penna dall'A. fin al C. toccando la cima dello ſgarramento, & calcagnuolo, le linee, che formano la rota di Poppa, ſi come l'infraſcritto diſegno ci moſtra.

Ruota di Poppa.



B.D. ſquadro palmi 13

B.Z. $\frac{1}{3}$ di palmo d'altezza della Carena.

I. Loco della Centa.

A. Draganti palmi $\frac{2}{3}$.

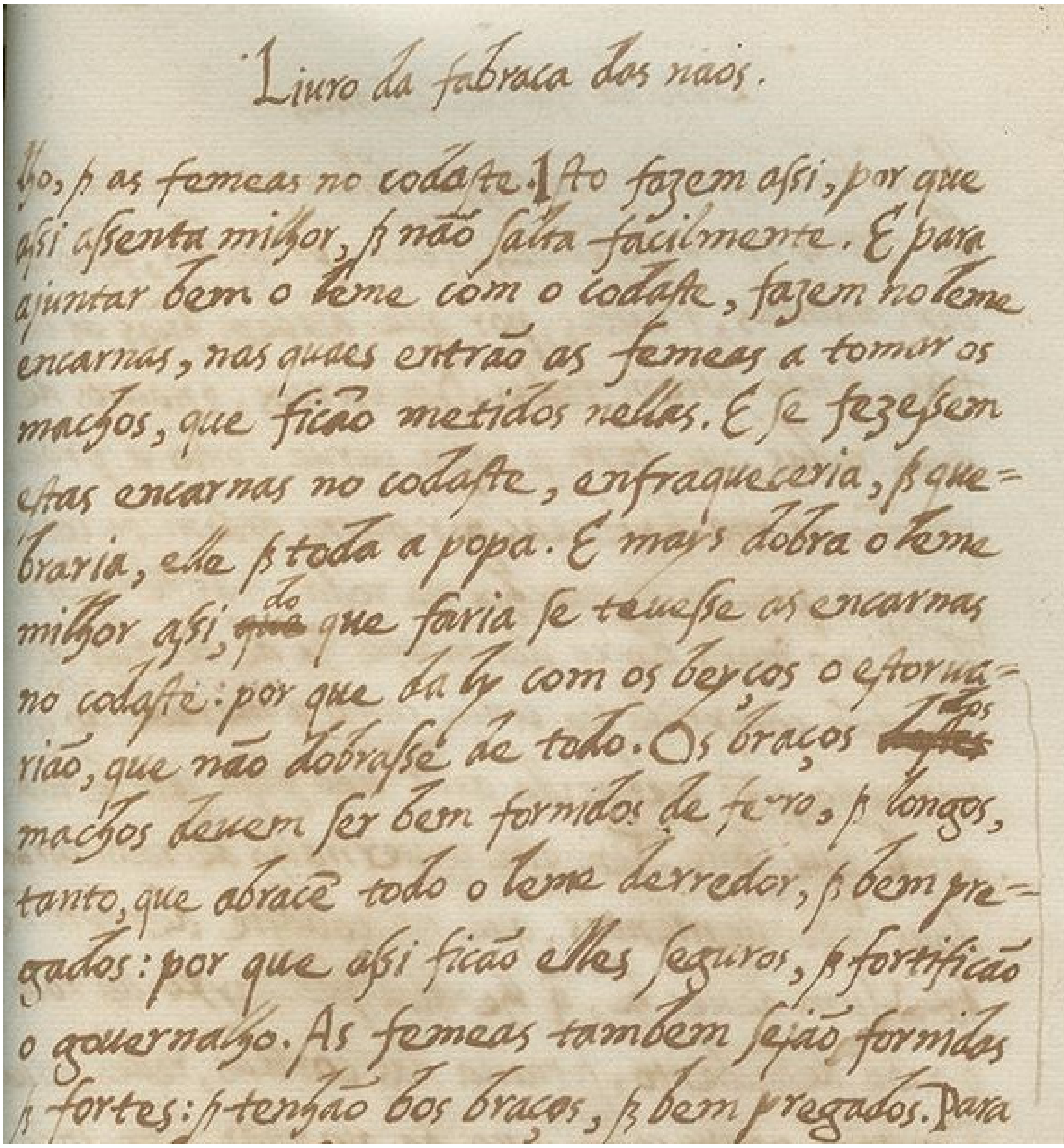
D.A. Palmi 3.

Y.I. Caparrone palmi 1.

B 2 B.E

Figure 4.4. Partial depiction of a page from 1601 Bartolomeo Crescenzio's *Nautica Mediterranea* (in Italian).

In the case of manuscripts, Figure 4.5 shows a partial image of a page from *Liuro da Fabrica das Naus*, written by Fr. Fernando Oliveira and dated to 1580. This is one of the earliest surviving Portuguese manuscripts. It includes dimensions of the main structural components of the ship along algorithms used in calculating measurements. Figure 4.6 depicts a page (text and illustration) of João Baptista Lavanha's *Livro Primeiro de Arquitectura Naval*, (c. 1600). The importance of this treatise lies in its accurate description of construction techniques and its detailed illustrations. Figure 4.7, offers a glimpse into 1616 Manoel Fernandes' *Livro de Traças de Carpintaria*. This work is divided into two main sections. The first lists dimensions of the primary structural components of a ship, while the second section provides a large collection of drawings depicting these components. Finally, Figure 4.8 depicts a folio of *Livro Nautico* (author unknown). This is a Portuguese collection of manuscripts bound in two volumes and dating between 1575 and 1625. The section devoted to ship construction offers detailed construction sequences.



Liuro da fabraca dos naos.

lho, & as femeas no codaste. Isto fazem assi, por que
assi assenta milhor, & não salta facilmente. E para
ajuntar bem o leme com o codaste, fazem no leme
encarnas, nas quaes entrão as femeas a tomar os
machos, que ficão metidos nellas. E se fezessem
estas encarnas no codaste, enfraqueceria, & que=
braria, elle & toda a popa. E mays dobra o leme
milhor assi, ~~que~~ do que faria se teuesse as encarnas
no codaste: por que da hy com os beyços o estorua=
rião, que não dobrasse de todo. Os braços dos
machos deuem ser bem fornidos de ferro, & longos,
tanto, que abrace todo o leme derredor, & bem pre=
gados: por que assi ficão elles seguros, & fortificão
o gouernalho. As femeas tambem sejão fornidas
& fortes: & tenhão bos braços, & bem pregados. Para

Figure 4.5. Partial depiction of 1580 Fernando Oliveira's *Liuro da Fabrica das Naus* (in Portuguese).

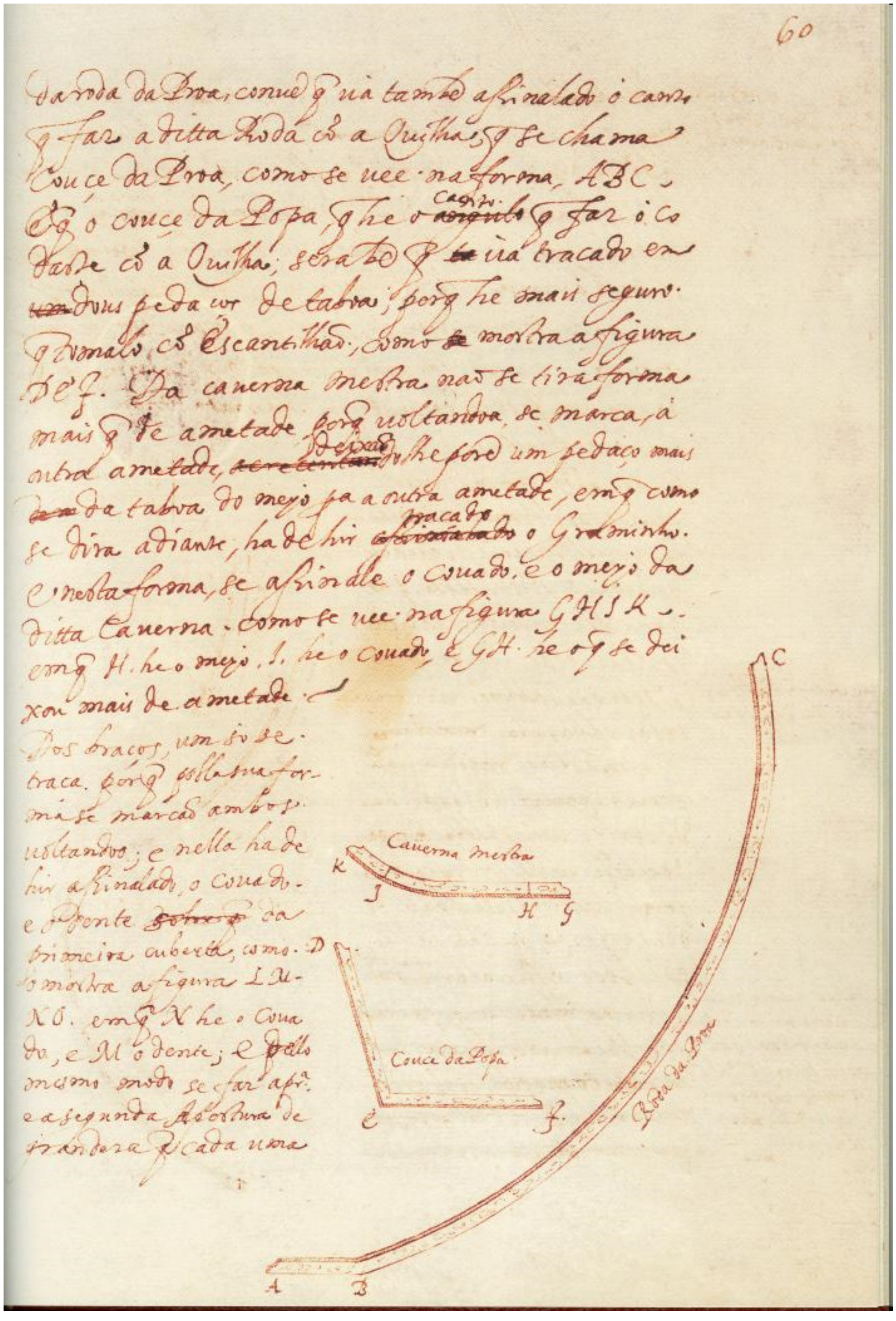
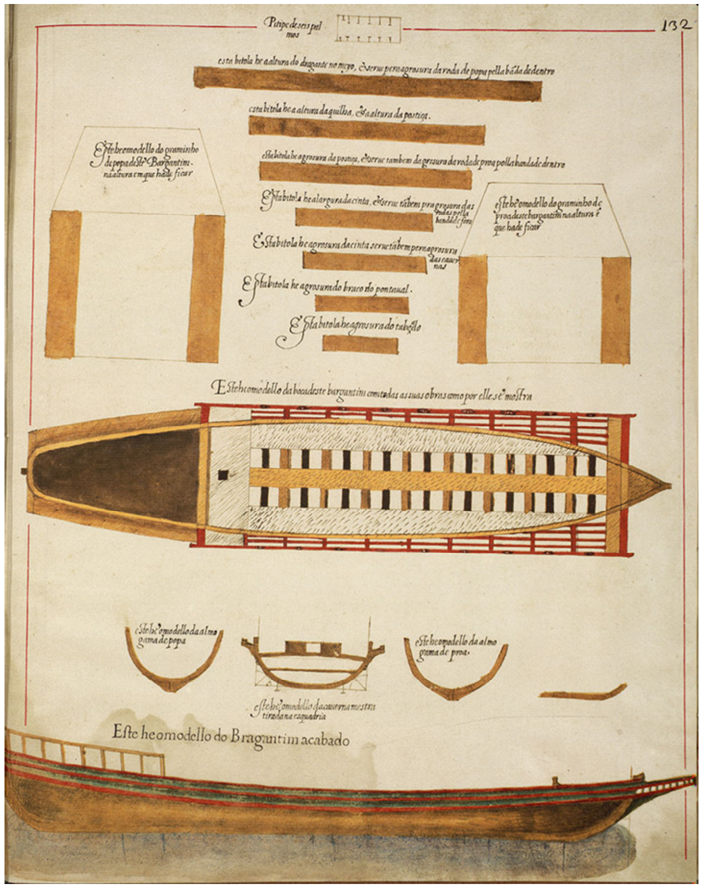
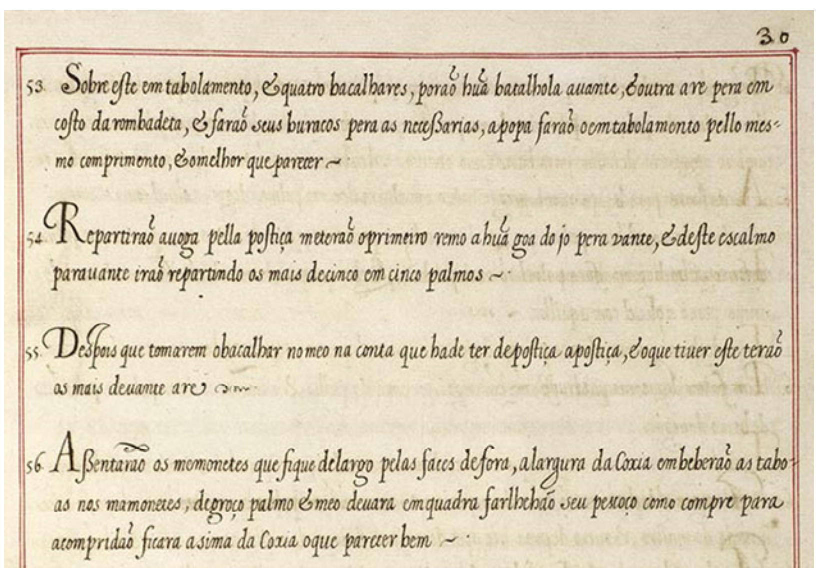


Figure 4.6. Partial depiction of a page comprising text and illustrations from Lavanha's *Livro Primeiro de Arquitectura Naval* (c. 1600).





30

53 Sobre este em tabolamento, & quatro bacalhares, porão hũa bacalhola avante, & outra a re pera em costo da rombadeta, & farão seus buracos pera as necessarias, a popa farão o em tabolamento pello mesmo comprimento, & o melhor que parecer ~

54 Repartirão a voga pella postiça meterão o primeiro remo a hũa goa do jo pera vante, & deste escalmo para avante irão repartindo os mais de cinco em cinco palmos ~

55 Despois que tomarem o bacalhar no meo na conta que hade ter de postiça a postiça, & o que tiver este terão os mais de vante a re ~

56 Assentarão os momonetes que fique de largo pelas faces de fora, a largura da Coxia embeberão as taboas nos mamonetes, de groço palmo & meo de vara em quadra farlhe hão seu pescoço como compre para a compridão ficara a sima da Coxia o que parecer bem ~

Figure 4.7. Depiction of an illustration and assembling instructions from 1616 Manoel Fernandes' *Livro de Traças de Carpintaria* (in Portuguese).

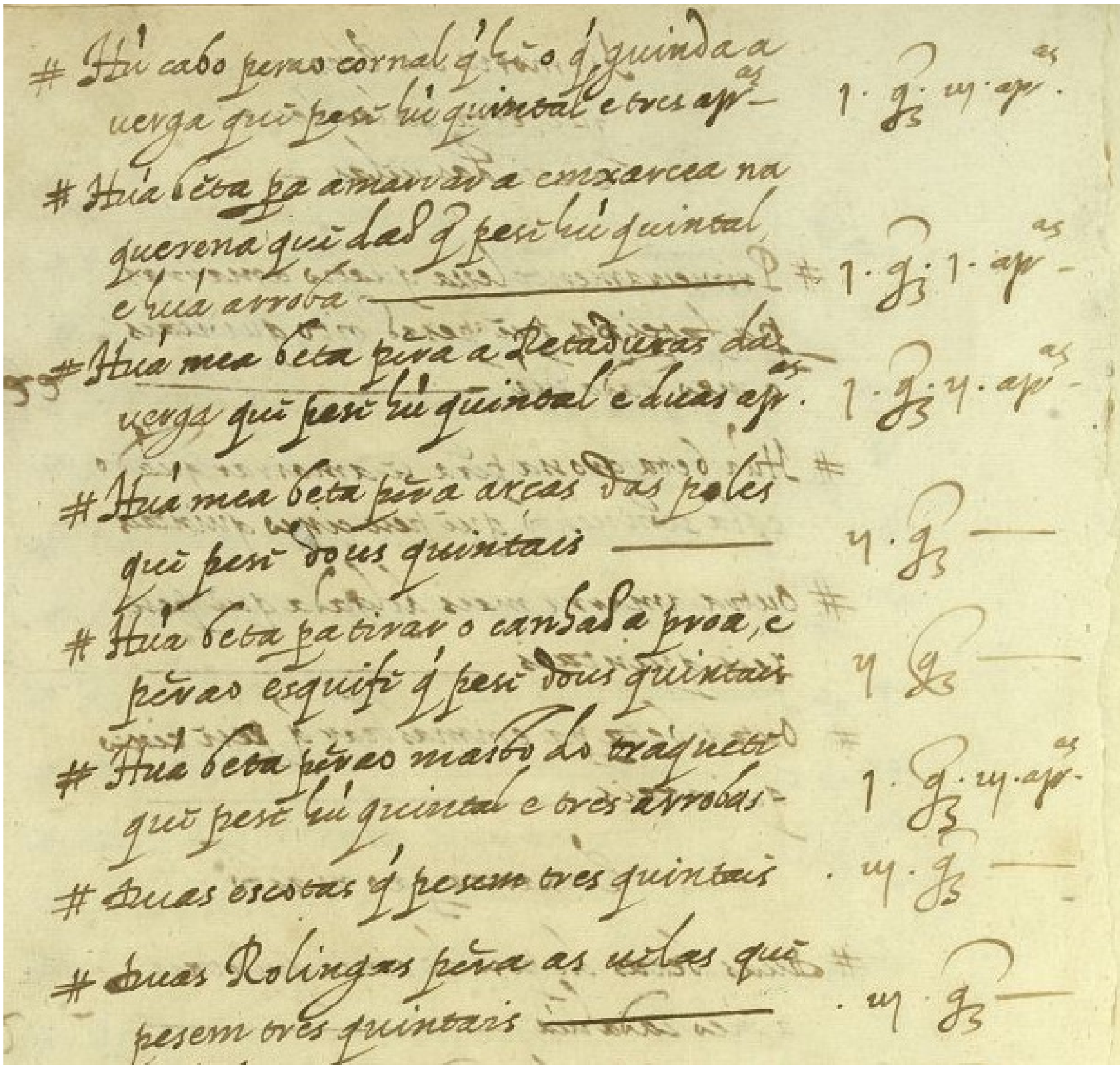


Figure 4.8. Partial depiction of a page in *Livro Nautico*, 1575-1625 (in Portuguese, author unknown).

From an early oral tradition, where descriptions and instructions were passed from masters to apprentices, shipbuilding practices evolved into a more formal discipline with the introduction of technical documents. In this transition, knowledgeable masters began to write down the instructions. The Age of Discovery, with Portuguese and Spanish sailors navigating to America, Africa, and India, was a driving force in this transformation. In the case of the Portuguese, the establishment of the trade routes to India required robust vessels to endure long voyages and weather conditions.

From the perspective of naval and seafaring dissemination, shipbuilding treatises are priceless sources for scholars working in ship reconstruction and studying the evolution of shipbuilding techniques. Moreover, the development of underwater archaeology in the last 50 years propitiated the growth of the archaeological data corpus, which can now be tested against the textual evidence pertaining to the conception and construction of these complex machines. Nautical Archaeology students, on the other hand, study ship treatises as part of their curriculum. Finally, for the general public, they are sources of historical and cultural contexts in which seafaring flourished.

The importance of shipbuilding treatises in the context of my dissertation is threefold. First, they provide numerous possibilities for linking textual descriptions, visual representations, and archaeological evidence; which are used for testing hypotheses in the reconstruction of an unknown composite physical object (ship). Second, their characteristics require particular indexing, segmentation, and visualization approaches for helping in disambiguating and understanding terms, concepts, and construction techniques across numerous naval traditions and their evolution over time. Third, they belong to a specific type of texts in a specialized domain for which little has been done in terms of creating a computer-based infrastructure.

The collection of treatises used in this dissertation was started with three Portuguese texts obtained with permission from the Academia de Marinha and National Library of Portugal. The collection grew both in quantity and diversity. At present it contains fifteen texts in Portuguese, French, Italian, Dutch, English, German, and Latin, spanning a period from the late 16th to the early 18th centuries. The list of treatises used in this dissertation can be found in appendix B.

4.3 Text Preparation

The main characteristics and contents of shipbuilding treatises were described in the previous paragraphs. The next step is the adoption of an encoding that enables text segmentation and mappings to multiple visual representations. In other words, texts have to be modeled in a way that makes their structure suitable to be represented, transformed, and manipulated in a computational framework. This model is important since the way text is represented on a computer affects they way it is used [DeRose et al. 1990].

Descriptive text encoding is one approach for text encoding. However, it seems to require in advance, knowledge about the relationships between segments. Referring to descriptive text encoding [Huitfeldt 1995] asserts that:

> There are several different kinds and uses of text encoding. The purpose of descriptive text encoding is not to prepare for some specific mode of presentation or analysis, but to represent as accurately as possible the textual information, the logical structure of the text, and the internal relationships between different text segments. (p. 236).

However, internal relationships of text segments is something (at least with the approach proposed in this dissertation) difficult to know in advance. The reason is due to the use of the glossary and the ontology for augmenting the original texts. In this context, relationships can be derived as more information is available, based on new discoveries reflected both in the glossary and the ontology.

A flexible and widely used model, hence the one I use, is adopting XML to encode the texts. This decision is primarily based on XML as open standard and W3C's recommendation. In addition and for practical purposes, XML-encoded texts can be easily transformed–using XSLT templates—allowing flexibility and scalability. The Renear-McGann debate previously quoted is relevant since it suggests two different theoretical approaches to texts: one hierarchical and the other dynamic.

Analyzing the contents and use of shipbuilding treatises, indicates they require a hybrid approach. Documents divided by sections and pages are useful for presenting the materials in a reader's interface, allowing parallel navigation (image-text) by page. It also provides a natural segmentation.

In full-text retrieval environments, text segmentation makes it possible to present query results in a more meaningful way, e.g. results divided by pages or paragraphs, or smaller text snippets. However, this "natural" and hierarchical approach limits the possibilities in which these documents can be used and manipulated. This observation coincides with Huitfeldt's critique of a hierarchical approach to texts [Huitfeldt 1995]:

> . . . why on earth should texts by all means be hierarchies? No doubt, there are many hierarchical structures, and no doubt this is important, but there are countless other relations between text elements which are worth while finding and investigating—overlap, substitution, discontinuity, parallel texts, cross references, etc. (p. 240).

Huitfeldt is not claiming that texts are not hierarchies at all. Rather, he suggests discovering new ways in which text elements are related; thus, enabling different representations. Ultimately, what I find in Huitfeldt's words is an encouragement for the adoption of unconventional approaches to texts. This unconventionalism is relevant to the textual representations of ships contained in the treatises.

For those treatises where transcriptions were available, a single XML file of the entire document was created. Page separation was indicated with a suitable tag and attribute value. The adopted encoding standard was Unicode (UTF-8). This encoding allowed handling words containing special characters in foreign languages. For each page, a reference to its corresponding image was inserted. A parser was created to separate the one-file document transcription into individual files. One file per page was created, and a unique identification number was assigned to each one of them. Figure 4.9 shows a partial segment of a transcription (one page) encoded in XML. The only divisions included are page and line. Similarly, the only reference included is a link to the corresponding image. The Document Type Definition (DTD) for this XML file can be found in Appendix D.

```
- <folio number="161">
    <graphic url="oliveira/f173.jpg" mimeType="image/jpg"/>
  - <l>
      <text>uela; e doutra feyção não gouernaraa. O gouernalho</text>
    </l>
  - <l>
      <text>de roda he hua tauoa, ou tauoas, a maneyra de</text>
    </l>
      .
      .
      .
      .
  - <l>
    - <text>
        porta, postas detras da popa, junto ao codaste; ou aa
      </text>
    </l>
  </folio>
```

Figure 4.9. A partial view of the XML-encoded version of a transcription.

4.4 Structuring Shipbuilding Treatises

The preliminary text preparation just described creates a basic document structure. As additional and more complex semantic structures were needed, limitations emerged. For example, a page-based segmentation is not suitable for cases where a technical description or an assembling instruction spans over various pages, see Figure 4.10. A construction sequence can also extend across more than one page (see Figure 4.11). Similarly, within the same page (or across various pages) it is not uncommon to find overlapping construction sequences based on ship components (see Figure 4.12). In addition, the encoding has to enable multiple transformations; functional, spatial, and sequential to name a few.

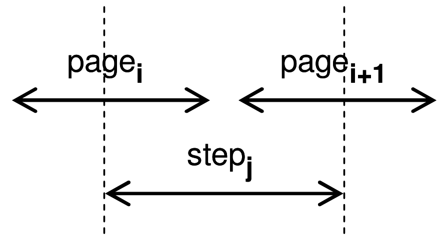


Figure 4.10. Diagram depicting an assembling instruction (step j) overlapping across two pages.

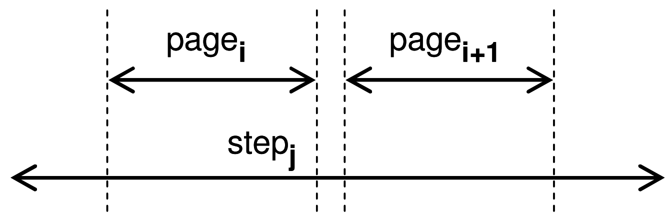


Figure 4.11. Diagram depicting a construction sequence (step j) expanding across more than two pages.

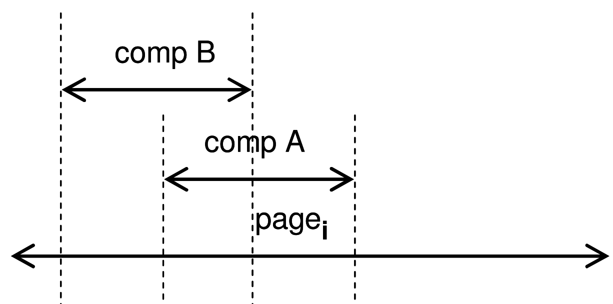


Figure 4.12. Diagram depicting two components (comp A and comp B) overlapping in the same page and spanning more than one page.

One of the early attempts to establish a standard for describing the structure and properties of documents was ODA (Open Document Architecture). With the goal of separating document components, the ODA model defined a separation between the logical structure within a document (chapters, paragraphs, etc.), the layout structure (pages, columns, etc.), and the actual contents of the document [Appelt et al. 1990]. In the ODA model, the logical and layout objects in a document are represented with the structural model [Appelt 1988]. Information pertaining to these objects was contained in attributes of a data structure called constituent.

Separating structure and layout from representation in a document is a feasible approach to tackle the shortcomings described in the previous sub-section. For instance, construction sequences and important terms or concepts derived from the glossary or the ontology, can be encoded regardless of their physical layout in the document. In this sense, an XML-encoded document can be seen as a hierarchy.

My approach in augmenting the semantics of the texts is based on the incorporation of two augmenting tools: a specialized glossary *glosShip* (described in Section 4.5), and a ship ontology *ontoShipDS* (described in Sections 6 and 7). In both cases the goal is to exploit the contents of the glossary and the ontology to highlight important segments of the texts to expand knowledge about terms; to enable linking of visual information from illustrations and photographs; and to allow term alignment in different languages—which is a crucial feature in the case of a multilingual environment.

Linking terms in the transcriptions with references in the glossary was achieved by matching both occurrences. However, instead of a one-to-one matching, an expanded matching algorithm was used. *Roles* of a term served as pivotal table, allowing the inclusion of synonyms and spellings (*roles* are described later in this section).

To achieve better matching results, I adapted Lucene's stemmer [The Apache Lucene Project 2009] to normalize words. If a normalized word from the text has a normalized entry in the glossary, the corresponding text string in the document is normalized and properly encoded (enclosed in a tag). Simultaneously, the index is updated with an instance for that word. Figure 4.13 (1) shows a segment of an XML-encoded file before being processed with *glosShip*. A transcription with generated hyperlinks to entries in *glosShip* is shown in Figure 4.13 (2).

Figure 4.14 is a partial depiction of two versions of the same Portuguese transcription (Oliveira's *Livro da Fabrica das Naos*). The text on the left is the pre-processed transcription—prior to the creation of glossary links. In the case of deletions/emendations, the original terms are depicted (below the red arrows: *que* and *destes*). The text on the right is the post-processed version. References to the glossary are shown in blue and underlined. For deletions/emendations, the emended text is shown (words in maroon, below the red arrow: *do* and *dos*). The red arrows in these examples are not part of the interface.

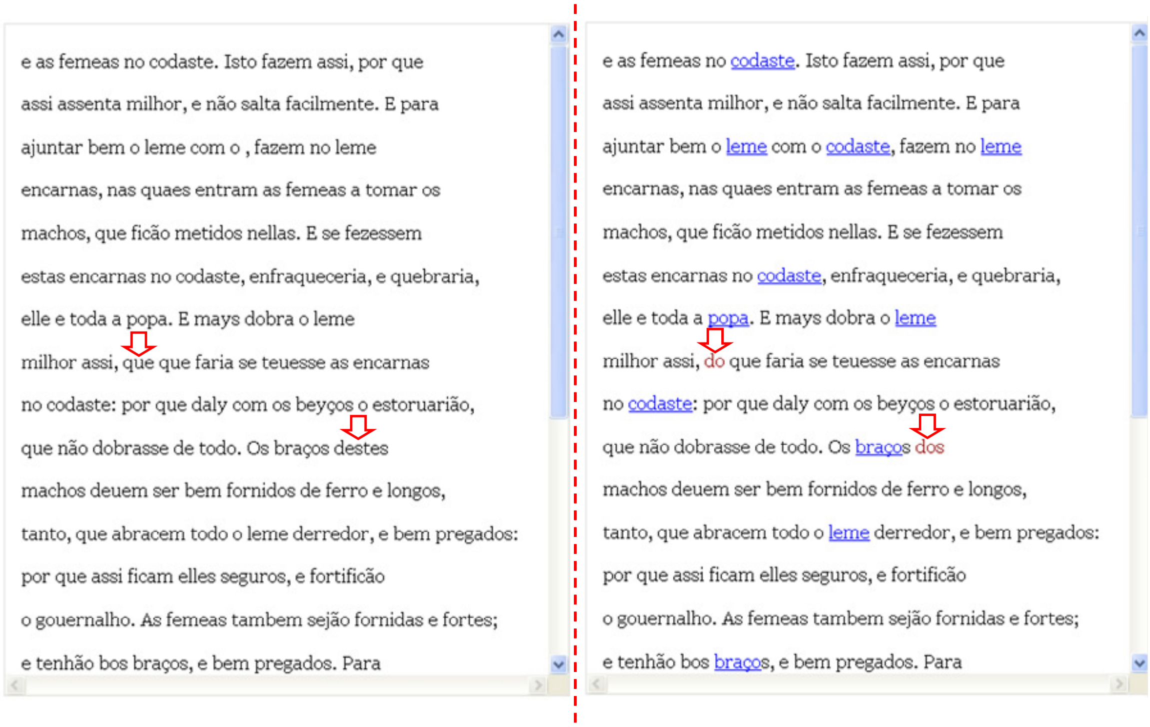
1


```
- <folio number="161">
    <graphic url="oliveira/f173.jpg" mimeType="image/jpg"/>
  - <l>
      <text>uela; e doutra feyção não gouernaraa. O gouernalho</text>
    </l>
  - <l>
      <text>de roda he hua tauoa, ou tauoas, a maneyra de</text>
    </l>
      ▪
      ▪
      ▪
  - <l>
    - <text>
        porta, postas detras da popa, junto ao codaste; ou aa
      </text>
    </l>
  </folio>
```


2


```
- <folio number="161">
    <graphic url="oliveira/f173.jpg" mimeType="image/jpg"/>
  - <l>
      <text>uela; e doutra feyção não gouernaraa. O gouernalho</text>
    </l>
  - <l>
      <text>de </text>
      <glossary id="138">roda</glossary>
      <text> he hua tauoa, ou tauoas, a maneyra de</text>
    </l>
      ▪
      ▪
      ▪
  - <l>
      <text>porta, postas detras da </text>
      <glossary id="139">popa</glossary>
      <text>, junto ao </text>
      <glossary id="143">codaste</glossary>
      <text>; ou aa</text>
    </l>
  </folio>
```


Figure 4.13. The preprocessed (1) and the post-processed (2) XML-encoded versions of the same page of a Portuguese treatise.


e as femeas no codaste. Isto fazem assi, por que
assi assenta milhor, e não salta facilmente. E para
ajuntar bem o leme com o , fazem no leme
encarnas, nas quaes entram as femeas a tomar os
machos, que fição metidos nellas. E se fezessem
estas encarnas no codaste, enfraqueceria, e quebraria,
elle e toda a popa. E mays dobra o leme
milhor assi, que que faria se teuesse as encarnas
no codaste: por que daly com os beyços o estoruarião,
que não dobrasse de todo. Os braços destes
machos deuem ser bem fornidos de ferro e longos,
tanto, que abracem todo o leme derredor, e bem pregados:
por que assi ficam elles seguros, e fortifição
o gouernalho. As femeas tambem sejão fornidas e fortes;
e tenhão bos braços, e bem pregados. Para

e as femeas no codaste. Isto fazem assi, por que
assi assenta milhor, e não salta facilmente. E para
ajuntar bem o leme com o codaste, fazem no leme
encarnas, nas quaes entram as femeas a tomar os
machos, que fição metidos nellas. E se fezessem
estas encarnas no codaste, enfraqueceria, e quebraria,
elle e toda a popa. E mays dobra o leme
milhor assi, do que faria se teuesse as encarnas
no codaste: por que daly com os beyços o estoruarião,
que não dobrasse de todo. Os braços dos
machos deuem ser bem fornidos de ferro e longos,
tanto, que abracem todo o leme derredor, e bem pregados:
por que assi ficam elles seguros, e fortifição
o gouernalho. As femeas tambem sejão fornidas e fortes;
e tenhão bos braços, e bem pregados. Para

Figure 4.14. Partial depiction of two versions of the same transcription.

The inclusion of links is a straightforward process, yet some interesting patterns emerge. Figure 4.15 shows a prototype of an interface for presenting text along with its corresponding image. The visual effect of colored words (those with references in the glossary) attracts the reader's attention. Skimming through the page reveals some repeated terms. An expert—someone knowing the meaning of those terms—can grasp a sense of the content of the page on reading the text. For a non-expert, this knowledge can be augmented by additional information associated to those words. For instance, the glossary allows categorization of terms. In the case shown here, the blue-underlined terms fit in two categories: structure and tools. This suggests that the topic of this page is most likely about structural components of the ship. In addition, from a spatial standpoint, it indicates that these descriptions correspond to the stern (back of the vessel). Words circled in red on the other hand, indicate corrections made to the manuscripts.

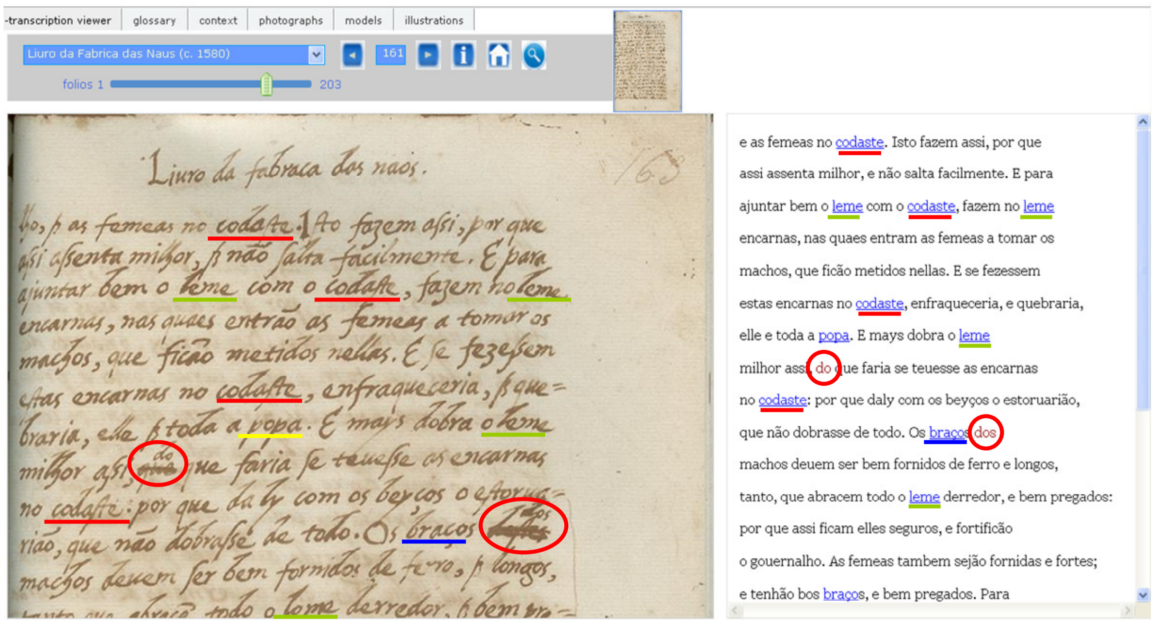


Figure 4.15. A screen shot of the reader's interface prototype. Colors indicate the correspondence of terms in the image and text.

Although this is a simple example, it captures the enhancements introduced by using colors; or as [Tufte 2001] asserts: "Tying color to information is as elementary and straightforward as color technique in art." Therefore, using colors to communicate more information is a viable approach. For instance different colors to identify categories—what Tufte refers as *using color to label* or *color as noun*—would offer more insight than simply coloring all terms with the same color. Referring to the role of color in information design, Tufte identifies four fundamental uses [Tufte 2001]:

> *to label* (color as noun), *to measure* (color as quantity), *to represent or imitate reality* (color as representation), and to *enliven or decorate* (color as beauty). (p. 81).

However, adding colors also brings disadvantages, of which I identify two. First, a term can belong to more than one category. Second, a page with numerous linked terms, the visual space can become cluttered and confusing. Despite these problems, the use of colors is a path worth exploring in future work.

In Section 5, I use an example of Google maps to illustrate the role of textual and visual representations simultaneously in problem solving (finding directions between two geographical locations). I argue that the contents of shipbuilding treatises can be used in a similar way for understanding construction sequences and relationships among ship components. In this context, my focus is on the relationships among different text segments, given the terms they include along with the categories and taxonomies they belong to.

4.5 *glosShip*: A Multilingual Glossary of Nautical Terms

The motivation in incorporating a multilingual glossary of nautical terms is based on discussions with nautical archaeologists about the contents of shipbuilding treatises (written in various languages). It was also influenced by the difficulties encountered when reading documents or listening to archaeologists using words and concepts difficult to understand. The English glossary included as appendix in [Steffy 1994] was used as starting point for *glosShip*. Terms from glossaries in other languages where later added. Although this glossary is comprehensive and detailed, it is constrained to one language, and suffers from the limitations imposed by printed media. In this context, a glossary is essential because nautical archaeology is a highly specialized domain where technical terms need to be explained for understanding their meaning and context.

We have previously discussed [Monroy et al. 2007a] a scalable infrastructure and a web-based interface for editing a multilingual glossary of nautical terms. Reading texts in a specialized domain can be difficult because of the expertise required in understanding words and concepts. A simple example to illustrate this point is the news media. Starting on the second part of 2007 until early 2009, the media was dominated by financial and economic news stories. At times, it was difficult for non-expert readers (like me) to understand the terminology of the news due to the lack of knowledge given the words used; words such as swaps, derivatives, buybacks, liquidity, toxic assets, and equity to name a few.

A solution to this problem is the use of thesauri and glossaries. A glossary provides a list of terms (glosses) in a domain, along with their definitions. For nautical archaeologists, language can be challenging when consulting shipbuilding treatises (mainly because they are written in different languages), and unknown terms and concepts are often employed. In fact, thesauri and glossaries have been used extensively in numerous digital libraries initiatives. The Perseus Project [Crane 2002] offers a good example of incorporating dictionaries in a classics collection.

Both glossaries and thesauri offer many advantages when used simultaneously with the contents of a digital collection. For instance, Arachne [Foertsch 2006] is an electronic repository (database) of the German Institute for Archaeology. Because archaeological objects are scattered across the world, Arachne offers multilingual access and thesaurus. The Getty Thesaurus of Geographical Names [Baca 2004] is another example of an external tool that can be incorporated into existing textual materials, enhancing searching and browsing, as well as disambiguating names, finding synonyms, and expressing categories for geographical locations.

4.6 The Architecture of *glosShip*

The infrastructure and the glossary created (the editorial part of the glossary is the work and contribution of numerous nautical archaeology experts) as part of this dissertation is named *glosShip*. This name derives from merging the words *gloss* and *ship*, and symbolizes the combination of words in various languages pertaining to ships. To enable scalability, this glossary is based on the concept of an entity with multiple properties and roles. Figure 4.16 depicts the correspondence between the conceptualization of the glossary and the editing interface, and is explained in the following paragraphs. This can be expressed as follows: an entity **E** (term) can have *n properties and m roles*; a property can be a characteristic, attribute, or feature (in the present case languages). Roles on the other hand, describe the functions related to the entity (in the present case the term itself, spellings, and synonyms). Therefore, each entity in the glossary can be represented as a matrix $\mathbf{E}_{\mathbf{n,m}}$, (expression 4.1) where $\boldsymbol{n}$ is the number of properties, and $\boldsymbol{m}$ the number of roles.

$$\mathbf{E} = \begin{pmatrix} \rho_{1,1}\ ,\ \rho_{1,2} \cdots \rho_{1,m} \\ \rho_{2,1}\ ,\ \rho_{2,2} \cdots \rho_{2,m} \\ \vdots \qquad \vdots \qquad \vdots \\ \rho_{n,1}\ ,\ \rho_{n,2} \cdots \rho_{n,m} \end{pmatrix} \tag{4.1}$$

In this case columns correspond to roles, and rows to properties. As stated earlier, every row in the matrix represents the translation of a given word into a different language. Each column on the other hand corresponds to the translation of a particular role of that term. The first column represents the base role—the term itself. The second column indicates a list of spellings of that term, and the third column a list of synonyms. Thus each cell $\rho_{i,j}$ in the matrix—with the exception of the first column—can be denoted as a vector of values:

$$\rho_{i,j} = \{v_1 \quad v_2 \quad \cdots \quad v_k\} \tag{4.2}$$

where $j > 1$ and $k \geq 0$.

Cells in the first column (base role) can have only one or zero values, and the base cell $\rho_{1,1}$ cannot be empty. Finally, a term can be associated to a set of taxa; this enables the creation of a taxonomy of terms.

Based on the architecture's formal definition previously described, the implementation of this multilingual glossary provides a multiple table-like interface where languages map the *properties*; and the term itself, spellings, and synonyms map *roles.* Also, entities (words) can have multiple definitions.

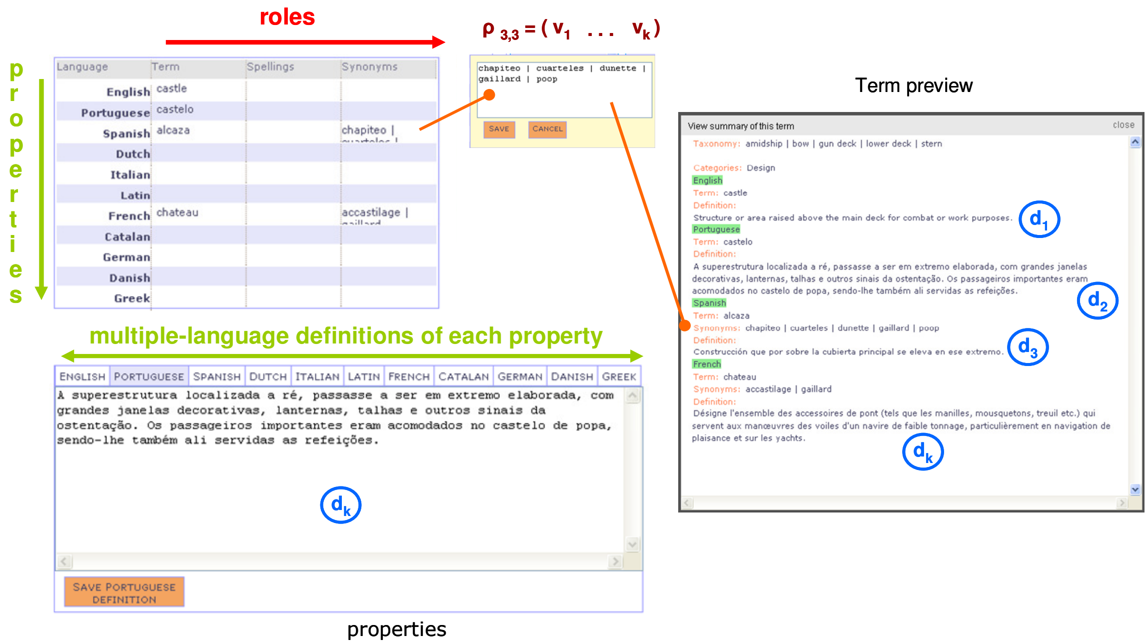


Figure 4.16. Correspondence between the conceptualization of the glossary and the editing interface.

4.7 *glosShip*'s Editing Interface

For practical purposes in the implementation of *glosShip*, characteristics, attributes, or features (such as languages) match *properties*. On the other hand, functions related to a term such as spellings and synonyms match *roles*. To expand the semantics of terms, glosses can be associated with taxa, allowing the creation of a taxonomy of terms, which in turn enables multiple text segmentation. Figure 4.17 shows the web-browser interface for editing terms. It allows remote access to scholars dispersed in different geographical locations. The interface is divided in 4 areas: 1) term list, 2) property/role editing area, 3) taxonomy/categories selector area, 4) definitions editing area. In the term list (area labeled 1) each column corresponds to a particular property—

languages in this case. Colors indicate the status of each term: a) red, a term without translation, b) blue, a term with translation, and c) green, a term with translations and definition. The term labels can be displayed in any given language. The property/role editing area allows entering information about the properties and roles of a term (area labeled 2). A pop up window allows editing the values (area labeled 6).

The taxonomy and categories options enable to associate taxa and categories from a multiple-selection list (area labeled 4). The definitions editing area allows entering the definition of terms in multiple languages (area labeled 3). Editors can also preview the information of a particular term (area labeled 5).

To certain extent this multilingual glossary shares some similarities with a traditional dictionary. However, the architecture extends the basic functions of a dictionary in two ways. First, it is not limited to a certain number of languages. This is important for this collection when incorporating new treatises from a particular naval tradition written in a language that was not originally included. Second, it allows multiple spellings, synonyms, and other *roles* that can be required. Spellings are very useful for archaeologists because technical terms in the manuscripts tend to have multiple spellings, as well as synonyms.

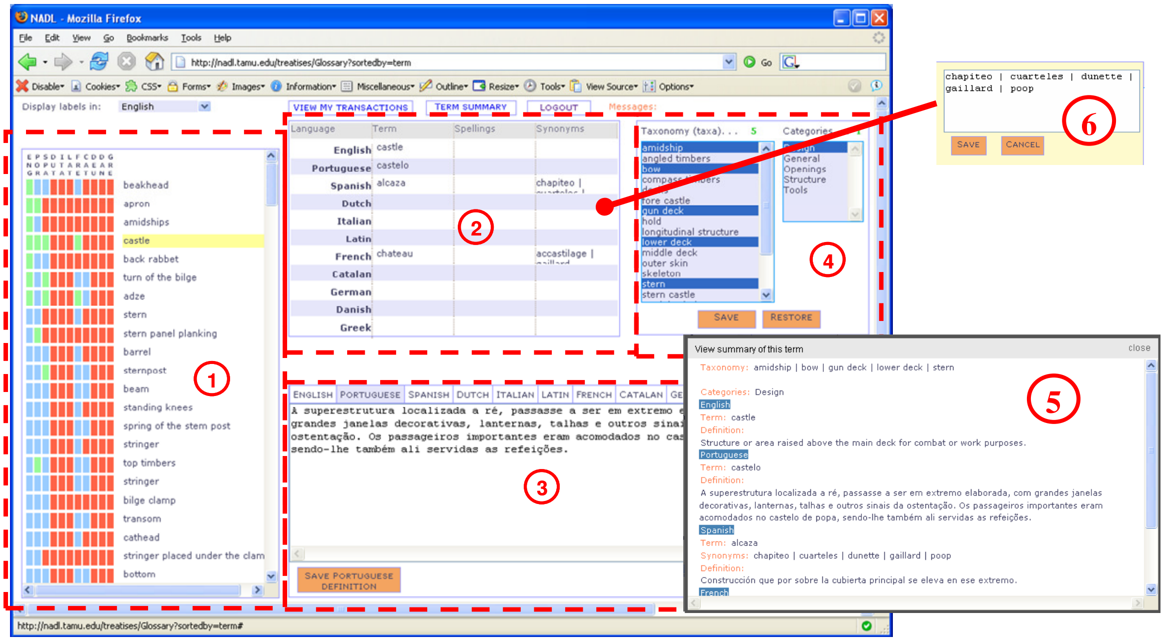


Figure 4.17. The web-based interface for editing terms and properties in various languages in *glosShip*.

5. VISUAL ABSTRACTIONS OF SHIPS

5.1 Introduction

In the preceding section I discussed a textual perspective for understanding ships and their components. I used the term *textual abstraction* because it is based on a written narrative that describes the characteristics of the various parts that make a ship, their properties, and how they are related. However, given the complexities of machines in general and ships in particular, relying solely on textual descriptions is not the best option.

A comprehensive analysis comparing the use of images and texts for describing physical composite objects would require a different study. In this section however, I highlight some points pertaining mostly to visual representations, although related to their textual counterparts (textual abstractions are discussed in Section 4). In the second part I explain my approach for augmenting visual representations of ships and how it fits into my proposed architecture.

Speaking about the properties of images and text [Molyneaux 1997] states:

> Many scholars think it is impossible (following Wittgenstein) for images to exist outside language. Artworks often mean little without the caption and the art history, the critic and the museum guide, to tell the viewer what to appreciate and how… At the same time, written language seems empty without imagery. (pp. 1-2).

Before starting this argument I would like to emphasize—in the case of ships—that textual descriptions and images are not mutually exclusive representations but rather complementary. A ship is after all a physical composite object. As such it can be represented as an aggregation of text snippets (this point was discussed in Section 4). Textual representations offer rhetorical tools, understood as a way of expression (similar to a dialog), hence the emphasis on the textual and linguistic characteristics. Ships represented through images, on the other hand offer semiotic tools, understood as a method for using and presenting symbols that visually convey a message or idea about an object (the ship).

Therefore Molyneaux's previous quote offers a compelling argument—to which I adhere—of the complementary nature of images and text. Therefore, it is not surprising when [Goggin 2004] talks about "visualizing the verbal, verbalizing the visual" when asserting that:

> In both a literal and figurative way, a rhetoric of the written word is visual, distinguishable from other forms of symbolic representation by the sense of sight. Both images and words on script, print or digital pages engage the eyes. When images and words appear together in one discursive space, they operate synergistically. (p. 88).

From the perspective of the computational framework I am proposing, this argument has important consequences. Mainly because the reconstruction of a damaged composite object—as is the case of ships—relies both on written and visual sources, especially considering that archaeological evidence is quite often subject to uncertainty. In short, textual information can complement visual representations and vice versa.

5.2 Visual Representations of Ships

I argue that visual representations can augment the understanding of a vessel and its components. For instance, understanding the results of a query in a collection of texts can be difficult if what is presented are the terms in isolation. In information retrieval and Internet search engines, an attempt to solve this problem is by presenting terms in context (lines or paragraphs). In nautical archaeology however, additional textual context does not necessarily imply a better way to grasp the meaning of ancient unknown or ambiguous technical terms and concepts. This happens especially when they correspond to visually rich concepts that require graphic depictions of the relationships among objects.

Therefore, a visual environment provides extra cues for understanding a term or concept. Viewing a term in association to other ship components can enhance cognition in two ways: a) the term itself, since related components give additional information, and b) the ship as a whole, since the aggregation of components helps to conceptualize how a vessel was built and what techniques and materials were used.

In this dissertation I consider three visual representations of ships: a) technical illustrations from the treatises, b) drawings, and c) photographs depicting timbers and ship fragments recovered from underwater archaeological excavations. Each one of these sources offers different perspectives in the conceptualization of a vessel, and is used in varied ways in the reconstruction of ships.

Illustrations in shipbuilding treatises visually represent the parts of the ship, and give a spatial context of the proportions of the vessel. Used in the reconstruction of ships and in the understanding of how they are built, visual abstractions are a great addition.

Working in the reconstruction of a Portuguese India Nau, and stressing the use of illustrations from shipbuilding treatises—Lavanha's *Livro Primeiro de Arquitectura Naval*—[Hazlett 2007] argues that the treatise ". . . is most useful for its accurate description of construction techniques and its detailed illustrations." I would like to highlight the fact that his emphasis is neither on the text nor on the illustrations alone, rather in the combination of both. Explaining the evolution in the contents of shipbuilding treatises, Hazzlet concludes:

> By the 18th century, shipwrights typically used lines drawings or models to visualize the vessel's shape before construction and to preserve that information after the vessel had been built. This was a fundamental departure from the simpler (and more conservative) non-graphic method used by 16th-century shipbuilders. (p. 42).

Graphic descriptions of vessels and their components in the treatises are quite diverse. The following three examples are taken from Lavanha's *Livro Primeiro de Arquitectura Naval*. In the first case (Figure 5.1), two individual components are depicted, along with the way they are assembled. The second example (Figure 5.2) shows the individual components that belong to a section of the vessel (the stern). The last case (Figure 5.3) shows a transversal view of a ship and all the components it comprises. Unlike the first two examples, this one gives a better overview of the proportions of the vessel (height and width).

Illustrations in the previous examples include labels; letters and numbers in figures 5.1 and 5.3, and letters and names in Portuguese (of ship elements) in Figure 5.2. The components within these illustrations (areas in the image) can be used as anchor points for linking texts. In other words, multiple textual narratives can be associated to these anchor points. This possibility enables not only text from the same treatise, but from other sources (even in various languages). This option highlights how written construction sequences can be enhanced when presented along with an illustration.

Conversely, an illustration (especially a technical one) is enriched with textual descriptions. Figure 5.4 shows a couple of ship parts with their corresponding written explanations.

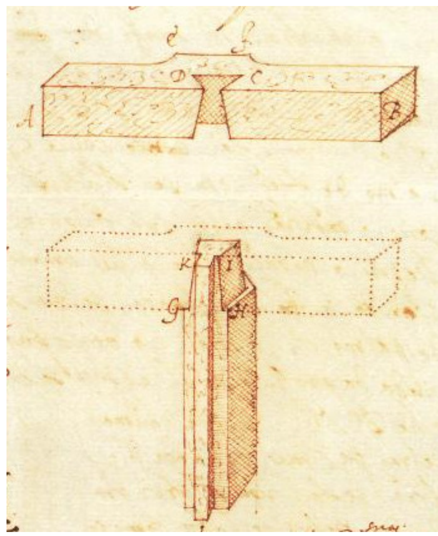


Figure 5.1. An illustration depicting two individual components and their assembling.

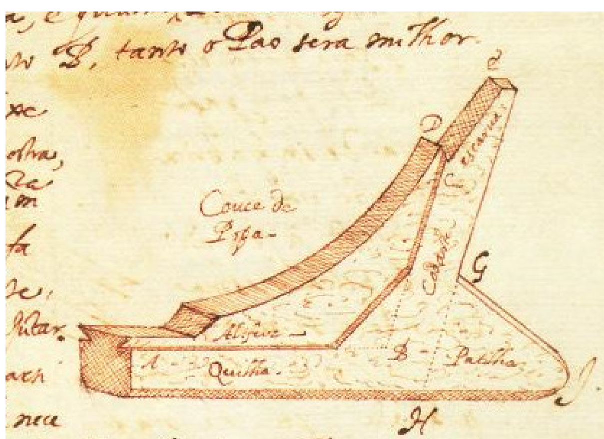


Figure 5.2. An illustration depicting a composite section of a ship and its parts.

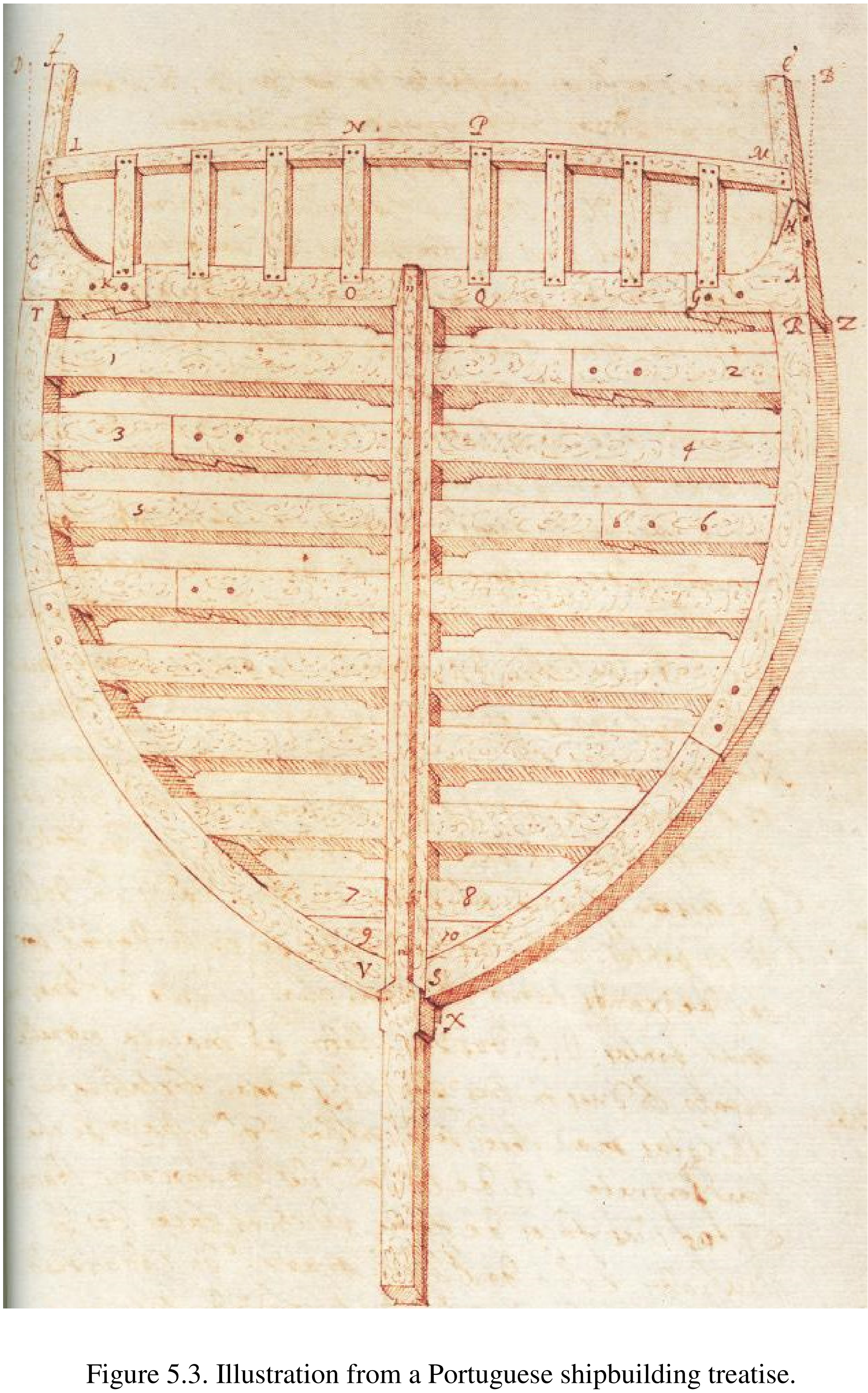


Figure 5.3. Illustration from a Portuguese shipbuilding treatise.

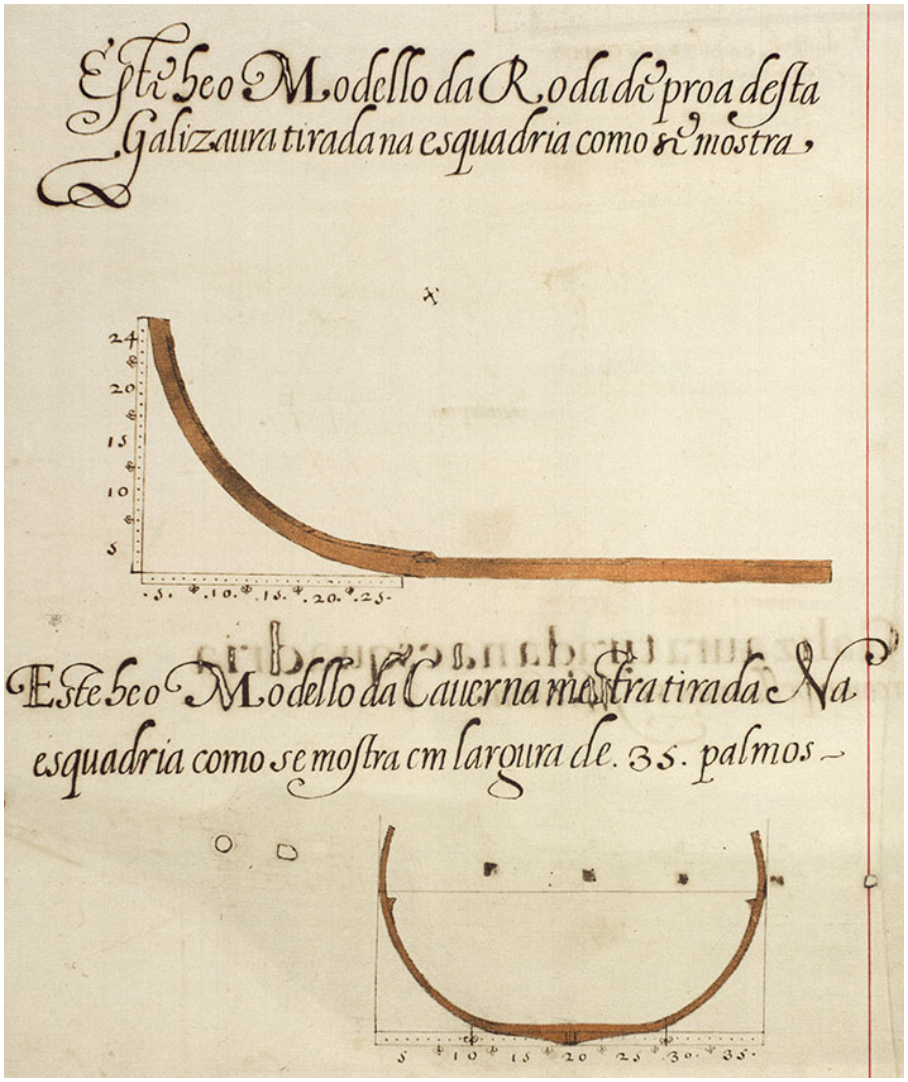


Figure 5.4. An illustration with its corresponding text (Manoel Fernandes' *Livro de Traças de Carpintaria).*

An example to illustrate the use of textual and visual abstractions in the same space for problem solving is the use of Google maps for finding directions. In Google maps, searching for directions from point A to B (although multiple destinations are allowed), requires entering the starting and ending addresses. A path-routing algorithm processes the request, the result is then presented visually and textually (Figure 5.5).

Graphic instructions are presented as a thick purple-colored line superimposed on the map, enabling the user to visually grasp distances, routes, and providing sense of direction and distances. Textual instructions, on the other hand are presented as a list of textual snippets. Each snippet indicates distance, estimated time, street names, and turns. Although both representations belong to the same reality, they convey the message in different ways. Text provides a step by step description, while the map offers a spatial overview of the route to be followed.

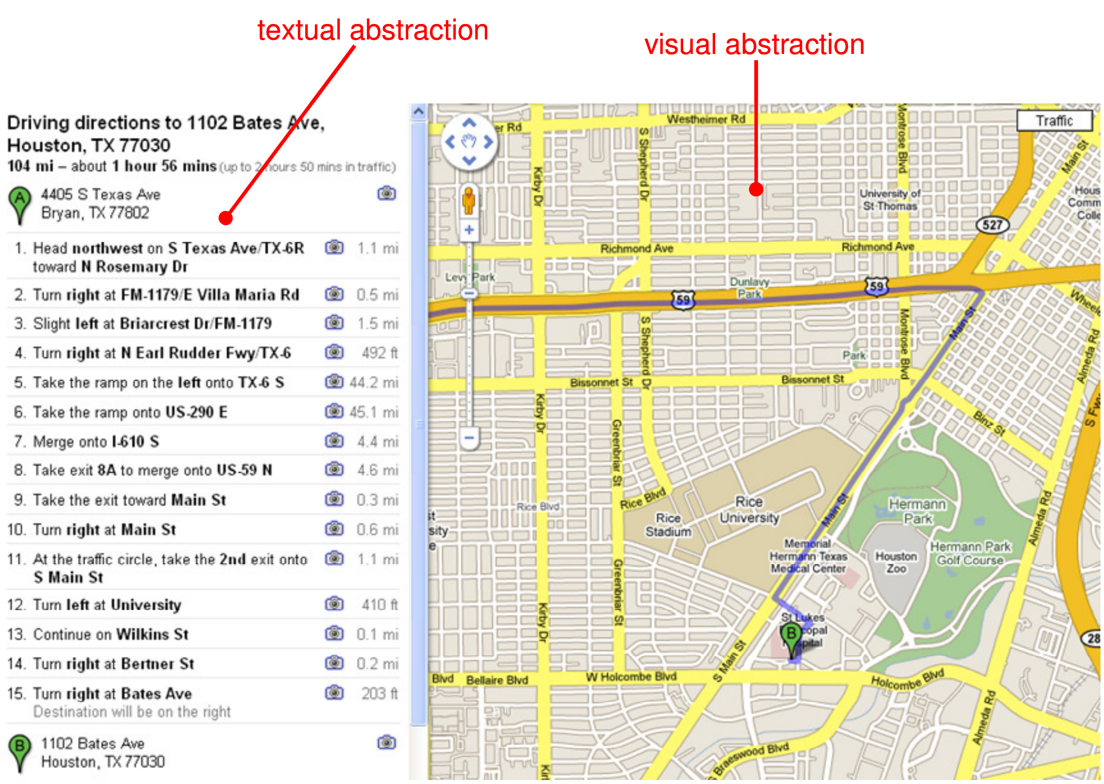


Figure 5.5. Partial image depicting driving directions in Google maps.

Visual representations of ships are not limited to illustrations in shipbuilding treatises. Throughout the excavation of a shipwreck, timbers and ship fragments are documented both textually as well as graphically. Textual documentation is comprised by notes taken by archaeologists at the excavation site and recorded in diver logs. Graphical documentation includes photographs and drawings about the ship components found.

Throughout the archaeological excavation cycle numerous photographs are taken that document the ship components discovered. Figure 5.6 depicts timbers from a shipwreck still partially assembled, offering unique information about construction techniques. As the excavation progresses and the recovered timbers and ship remains undergo various conservation procedures, new photographs are generated, revealing patterns and details that were previously hidden. Figure 5.7 shows a segment of a keel after being excavated. Note the holes and wood cuts on the timber. Diameter of the holes can indicate fastening methods as well as assembling patterns. Figure 5.8 shows a fragment of a timber with an empty area suggesting a nail was once use there (1). In this case, even if fastening objects were not recovered, wood marks can help to hypothesize the dimensions and types of fastening objects employed. Similarly, the number of holes, their diameters and placement can suggest the types of timbers used even if none were recovered.

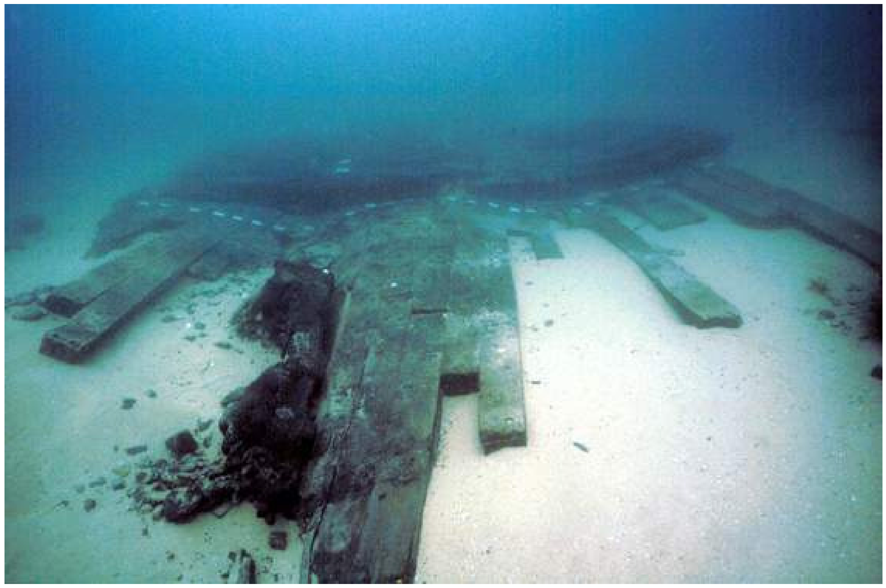


Figure 5.6. Ship remains (still assembled) from a Portuguese vessel.

Figure 5.7. Segment of the keel of an excavated ship showing some holes and cuts.

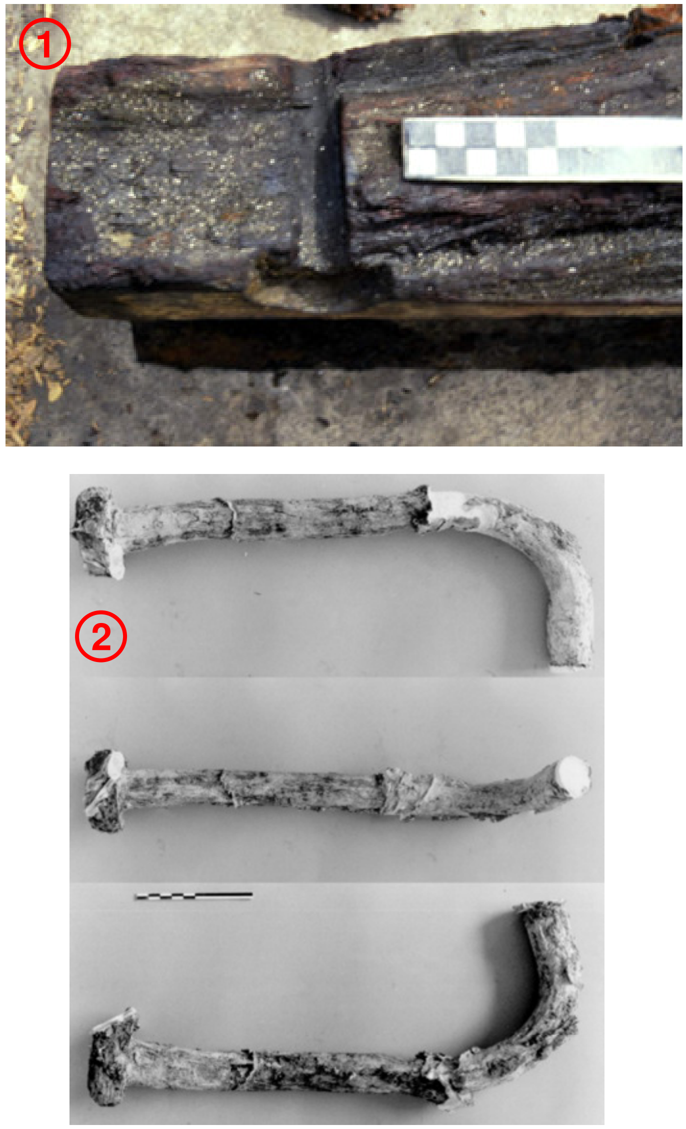


Figure 5.8. Photographs depicting part of a ship's timber with an empty area for a nail (1) and a few fastening pieces (2).

5.3 Incorporating Visual Representations into the Architecture

Although textual descriptions of ships and their components allow for more detailed explanations, visual representations on the other hand, provide affordances that their textual counterparts cannot. Proportions and associations among ship components are among the two most important characteristics—but not the only ones—that illustrations offer. Thus, it is not a surprise that shipbuilding treatises quite often include illustrations.

The integration of textual, visual, and archaeological sources in rebuilding a ship is highlighted in Hazzlet's reconstruction of a 16^{th}-century Portuguese Indiaman. The role of the texts in the reconstruction leads the author to name it a "textual excavation." Speaking about the creation of a 3D computer model, he first stresses the difficulty in understanding specialized ancient terms and techniques in a different language [Hazlett 2007]:

> Building the model has forced me to face not only the limits of our understanding of Portuguese shipbuilding nomenclature and methods... (p. 3).

When explaining the use of supporting sources such as written instructions from the primary treatise *Livro Nautico*, it becomes clear that they were meticulously consulted. Instead of being checked in isolation, they were complemented with data from other treatises. In addition, textual resources were complemented with visual materials and archaeological information:

> …using the scantling lists and instructions in the *Livro Nautico* line-by-line, supplemented with relevant data from other treatises as well as iconographic and archaeological data. (p. 5).

The process previously described sheds light into the complexities involved in the combination of diverse resources.

The methodology I propose for integrating illustrations and images into the framework requires first that a unique identification number be assigned to each image.

An image tagging interface *tagShip* (Figure 5.9) allows associating areas and points to the images. If an image is tagged with an area, a tuple in the following format is inserted in the database: (imageID,coordinates(x,y,w,h)). Conversely, if points were attached to the image, a tuple with the following format is inserted in the database: (imageId,coordinatesPoints(x1,y1,x2,y2,...xn,yn)).

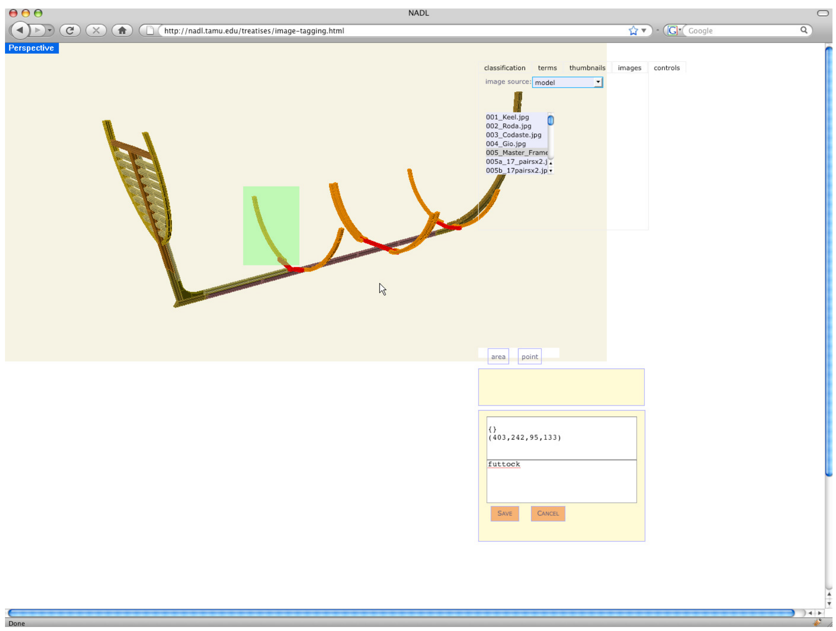


Figure 5.9. A screen shot of *tagShip*, the image tagging interface.

In order to associate texts to the illustrations, the user selects a term from the list provided. This information is then stored in the database. When a user is interested in images related to a particular term, the query fetches all entries in the table of illustrations with references to the term. Because the glossary includes categories and taxonomy, the searches can be expanded. Similarly, searching for a category returns terms that belong to the selected category.

The first layer is then at word level. A second layer of augmentation is the one that allows segments of the texts to be associated with components and areas in the illustration. This is an important architectural property because it allows textual segmentation based on construction sequences.

Illustrations in shipbuilding treatises correspond to a three-dimensional reality depicted in a two-dimensional space. One of the goals in using these illustrations for comparing and understanding shipbuilding techniques is to visualize various construction methods simultaneously. Scholars are interested—among other things—in finding how timbers or wooden fragments fit into the whole ship. Presenting and augmenting complex information and data in a two-dimensional space fits into what [Tufte 2001] refers to as "escaping flatland." (p. 12).

Images on the other hand complement textual representations. Linking and presenting these sources in a visual interface requires efficient indexing and suitable encoding, this in turns allows flexible retrieval for enabling discovery.

Various examples of the use of illustrations for helping understanding can be found in [Kirsh 2002]. Of those examples, two are relevant to ship reconstruction. One depicts the cardio-vascular system, and fits into the category of an illustration depicting parts and assemblies. In this case, Kirsh comparing the properties of a photograph and an illustration of the cardio-vascular system, states:

> The picture is cluttered and requires an experienced eye to identify parts. Even when major parts are labeled in the picture, it is still difficult to determine their relevant structure and their boundaries. (p. 2).

Pictures such as photographs of timbers and ship fragments recovered from an archaeological excavation give a more realistic representation of the physical entity, similar to the cardio-vascular system in this example. In the case of the illustration on the other hand, he adds:

> … we see that the key structural entities are shown in outline and simplified. Moreover, parts that are important elements of the system but virtually impossible to see in real life are enlarged and inserted into the illustration to allow viewers to see the system as a functioning whole. (p. 2).

The second example explains one function of illustrations; that is, to provide a visual depiction of written texts. It uses a story taken from a children's book. Although this is a simple example and its genre differs from a technical illustration pertaining to a ship, it points out to the effects of textual and visual representations of a given object or theme. Kirsh adds,

> There is no requirement to assume that the 'real' situation referred to in the text is identical to the depiction. The illustration may exaggerate features, it may violate certain naturalistic assumptions for comic or other effect, or it may hide from view details of real situations. At the same time, an illustration also typically adds details that are unmentioned in the text. (p. 2).

Sequences of ship construction are list of instructions—similar to the ones in the Google map example discussed earlier in this section. They describe the components of the vessel and how they are assembled. Similar to the maps, illustrations in shipbuilding treatises provide a visual abstraction of the components and their assembly. Because ships can be seen as machines, their illustrations fit into the categories of technical illustrations.

5.4 Technical Illustrations in Nautical Archaeology

Bethune defines technical illustrations as “the pictorial representation of technical information” [Bethune 1983], and suggests four categories of technical illustrations: a) simple drawings, b) exploded drawings, c) maintenance or assembly instructions, and d) cutaways. Despite that these illustrations belong to modern machines, they share certain similarities with ancient objects such as the ships described in the treatises. Since this kind of illustrations pertain to a very specialized domain, their understanding is often difficult, as Bethune writes:

> Technical information is so specialized that only persons trained in the particular area can clearly understand it. Each technical area has its own symbols, language, and conventions that, to the untrained person, are meaningless. (p. 1).

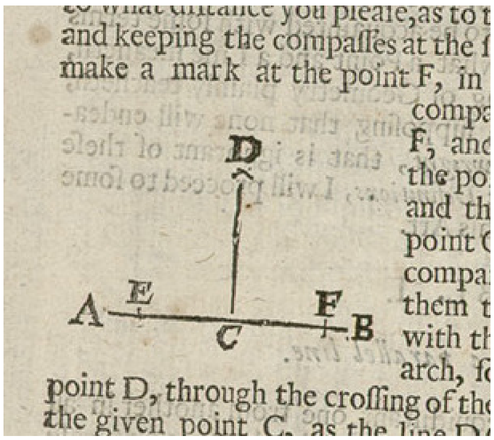
Conceptually, this is a top-down or overview-detail model because it offers a general view—complete or partial—of the system (ship), and its various sub-components. Using *ontoShipDS* it is possible to describe both the parts in the illustrations and how they are related (see Sections 6 and 7). *ontoShipDS* can also be used as a bridge to link the illustrations and the textual narrative included in the treatise. In a more ambitious scenario, it can be used to contrast the textual and visual abstractions of a ship in two or more treatises.

But Nautical Archaeology also provides a bottom-up or detail-overview model. This is based on the fact that the reconstruction of a ship starts from fragments, timbers, and ship parts recovered from the excavation. Scholars then analyze them figuring out how they were assembled and what the final ship looked like. From the information theory perspective, the practice of ship reconstruction offers interesting opportunities. For example, new discoveries might require the revision of the ontology. Or a more appealing solution would be to create a new ontology that can be compared with the existing one.

A few technical illustrations are shown in the following figures. Figure 5.10 depicts a couple of geometric drawings from Bushnell's treatise (in English). Figure 5.11 presents a list of components used in various parts of a vessel (from Bartolomeu's treatise). Figure 5.12 depicts the progression of timbers throughout the construction process: from trees, wood cuts, their placement in the ship, and dimensions. Figure 5.13 includes two small timbers within the text of a treatise. Both figures 5.12 and 5.13 are from the same French treatise.

A few examples from various Portuguese treatises can be seen in figures 5.14 to 5.18. Fernandez (1616) shows front views of the decks of a vessel along with dimensions (Figure 5.14). Figure 5.15 shows cuts and dimensions in various timbers of a ship. From Oliveira's manuscript, various geometric curves and lines are shown in figures 5.16 and 5.17. Lavanha illustrates a transversal view of the structural components of a ship, including floor frames and futtocks. Letters in the diagram correspond to geometric descriptions in the text (Figure 5.18). Figure 5.19 from the same treatise shows the floor frame attached to the keel.

From the Dutch shipbuilding tradition, Figure 5.20 shows an illustration from Van Ijk's treatise (1691). Geometric curves and lines are depicted along with timber cuts and shapes. Although this is a small sample of illustrations from the treatises, it offers a good idea of how a ship can be represented as a visual abstraction.



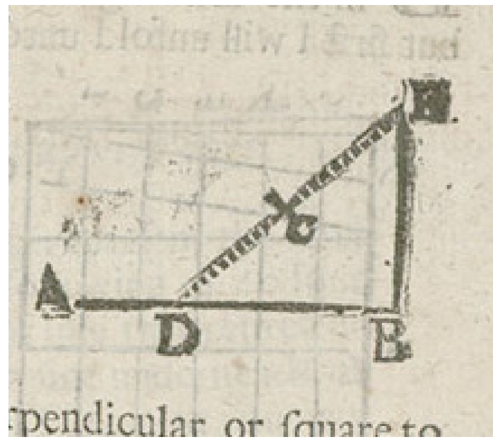


Figure 5.10. Partial depiction of two geometric illustrations in Bushnell's.

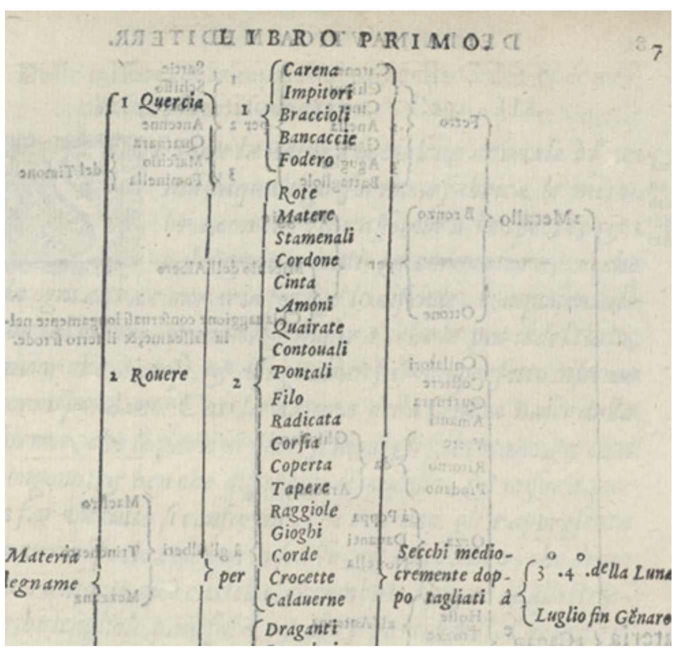


Figure 5.11. A taxonomy of wood species to be used in various parts of the ship. (Bartolomeu's treatise).

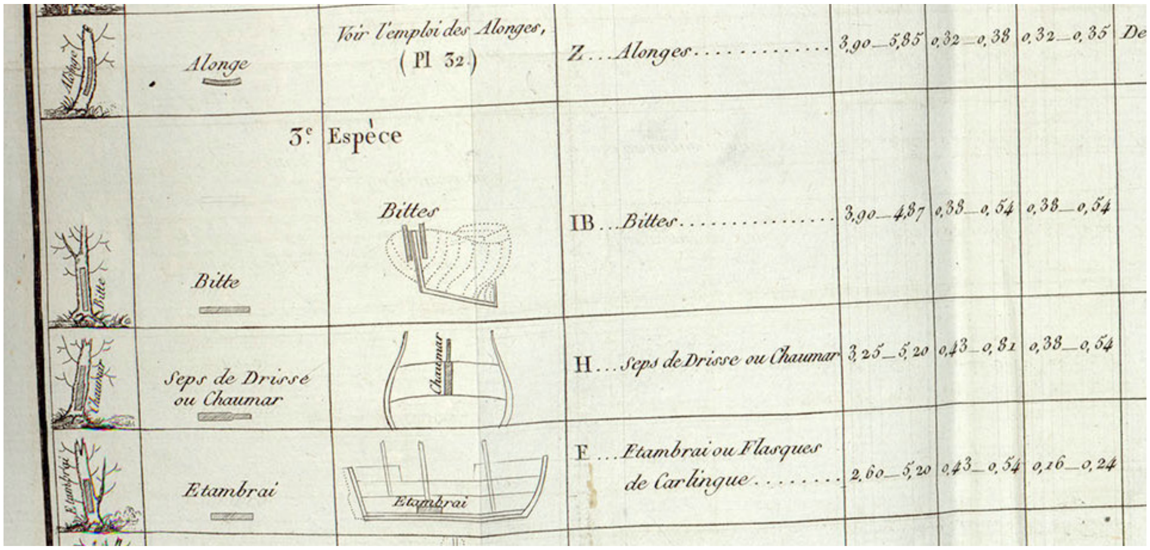


Figure 5.12. Illustration from the French treatise *Des Bois Propres Au Service des Arsenaux* (1813).

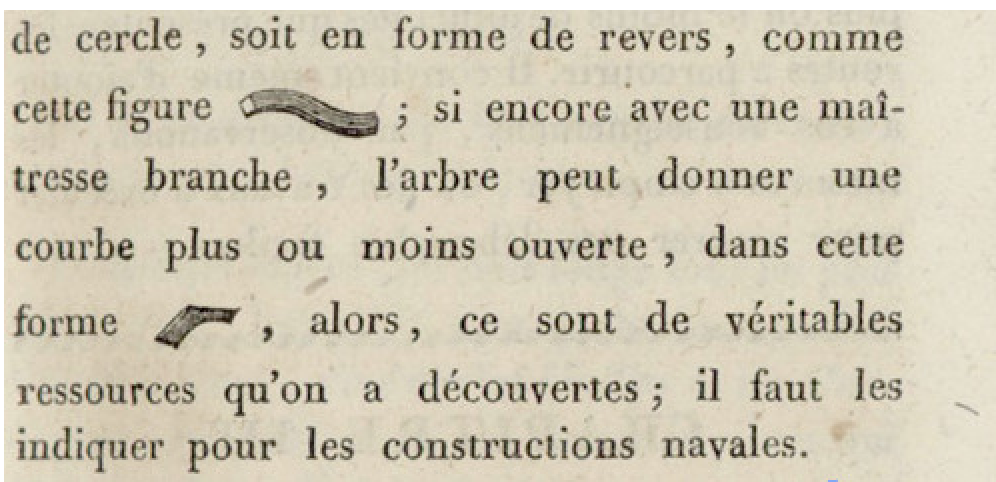

de cercle, soit en forme de revers, comme cette figure ; si encore avec une maîtresse branche, l'arbre peut donner une courbe plus ou moins ouverte, dans cette forme , alors, ce sont de véritables ressources qu'on a découvertes; il faut les indiquer pour les constructions navales.

Figure 5.13. Illustration and text fragment from the French treatise *Des Bois Propres Au Service des Arsenaux* (1813).

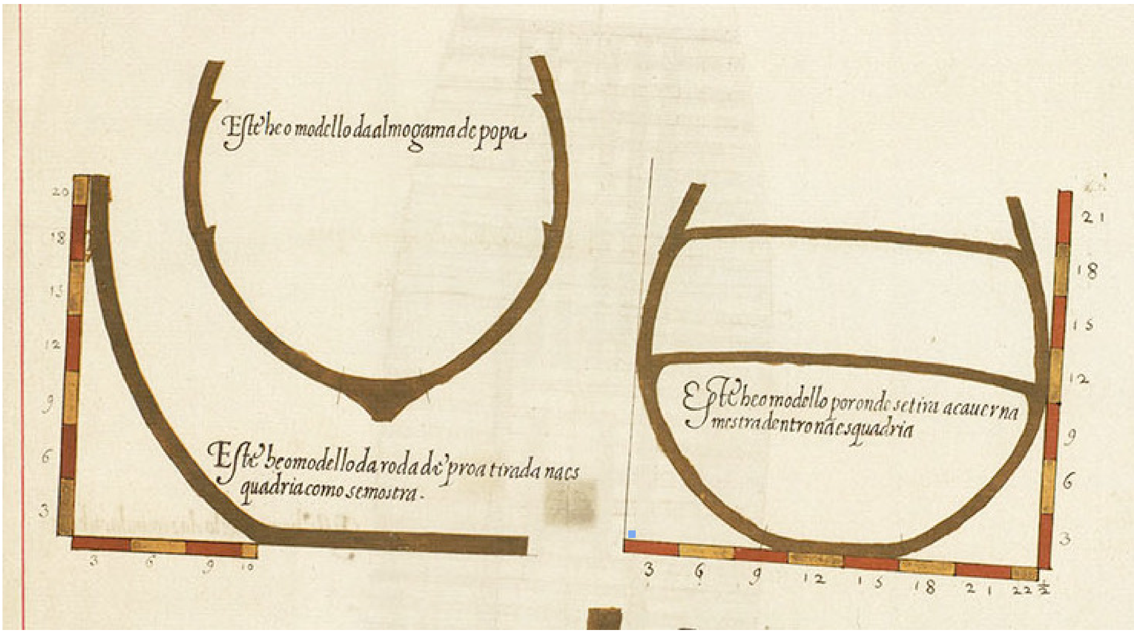


Figure 5.14. Illustrations and measurements from a Portuguese shipbuilding treatise (Fernandez 1616).

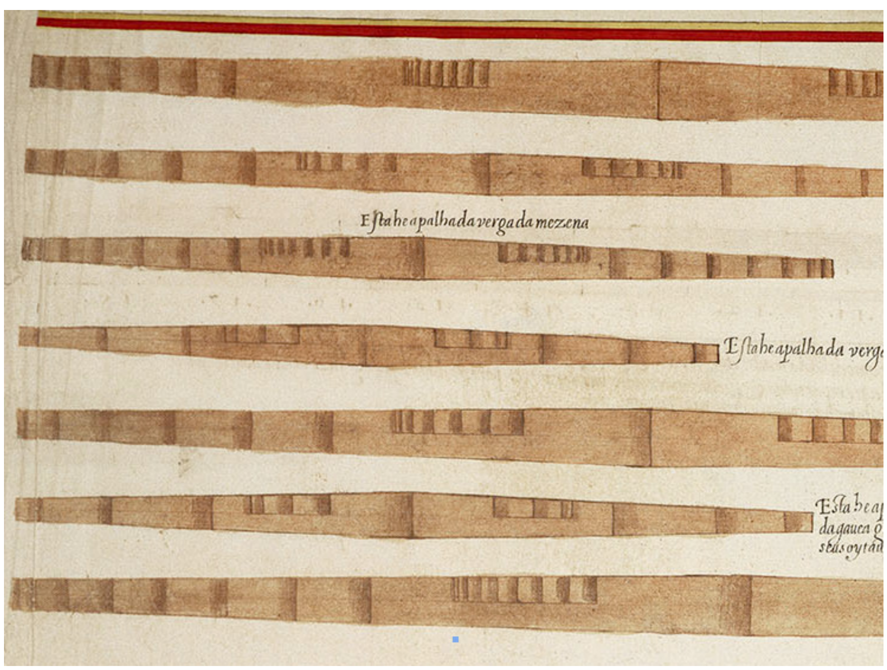


Figure 5.15. Technical illustrations from a Portuguese shipbuilding treatise (Fernandez 1616).

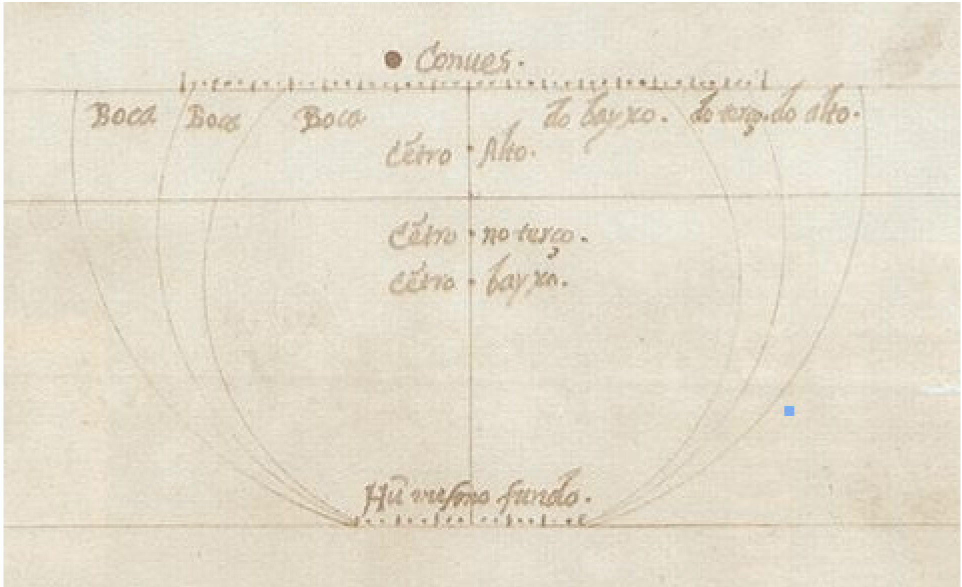


Figure 5.16. Illustration showing a transversal view of a vessel in Oliveira's *Liuro da Fabrica das Naus*.

Figure 5.17. Illustration showing a longitudinal view of a vessel in Oliveira's *Liuro da Fabrica das Naus*.

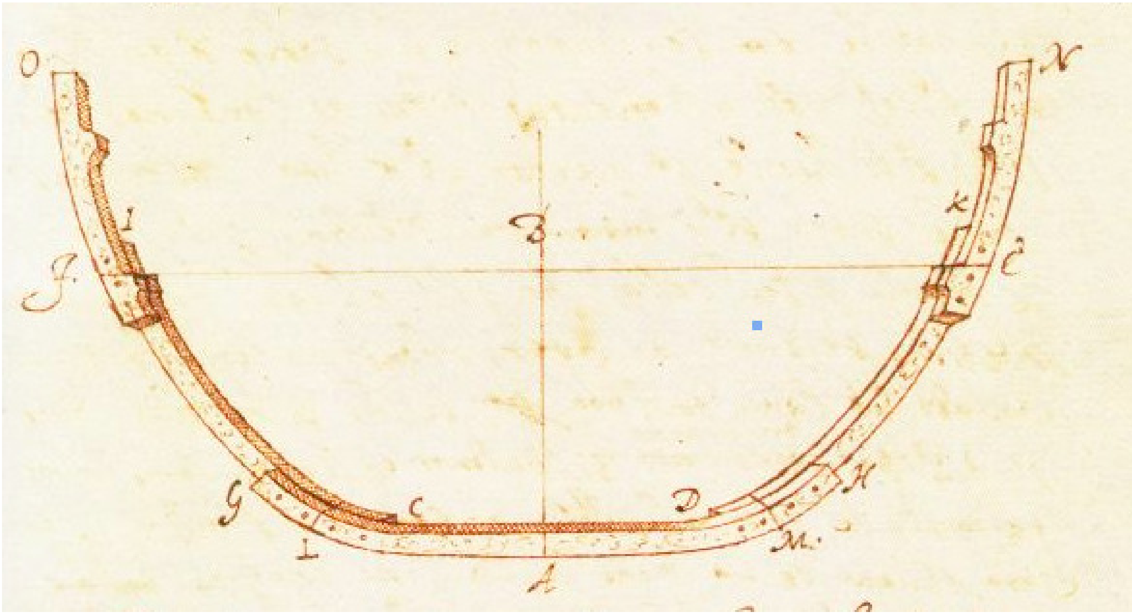


Figure 5.18. Lavanha’s depiction of a transversal structural part of a ship
(c. 1600).

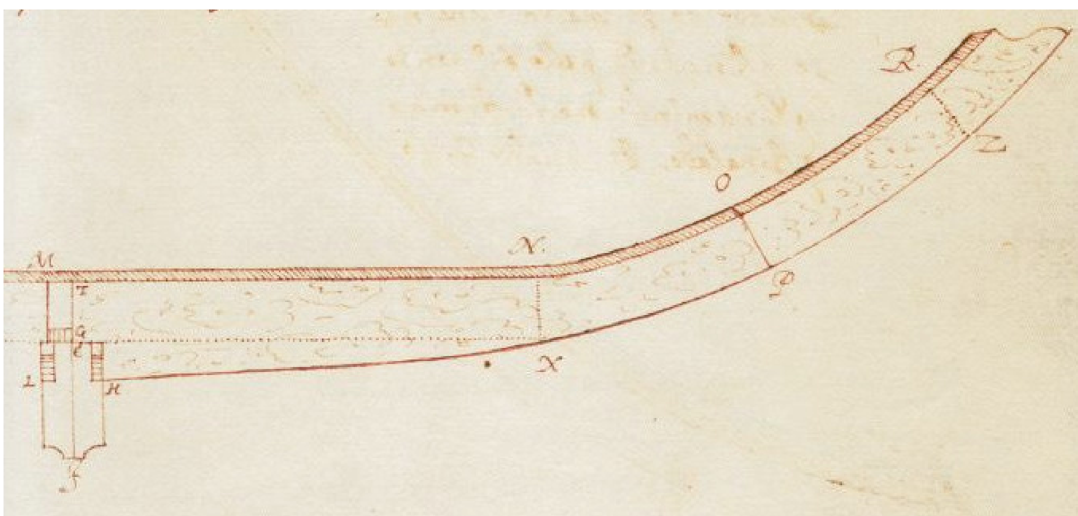


Figure 5.19. Floor frame attached to the keel. Illustration from Lavanha (c.
1600).

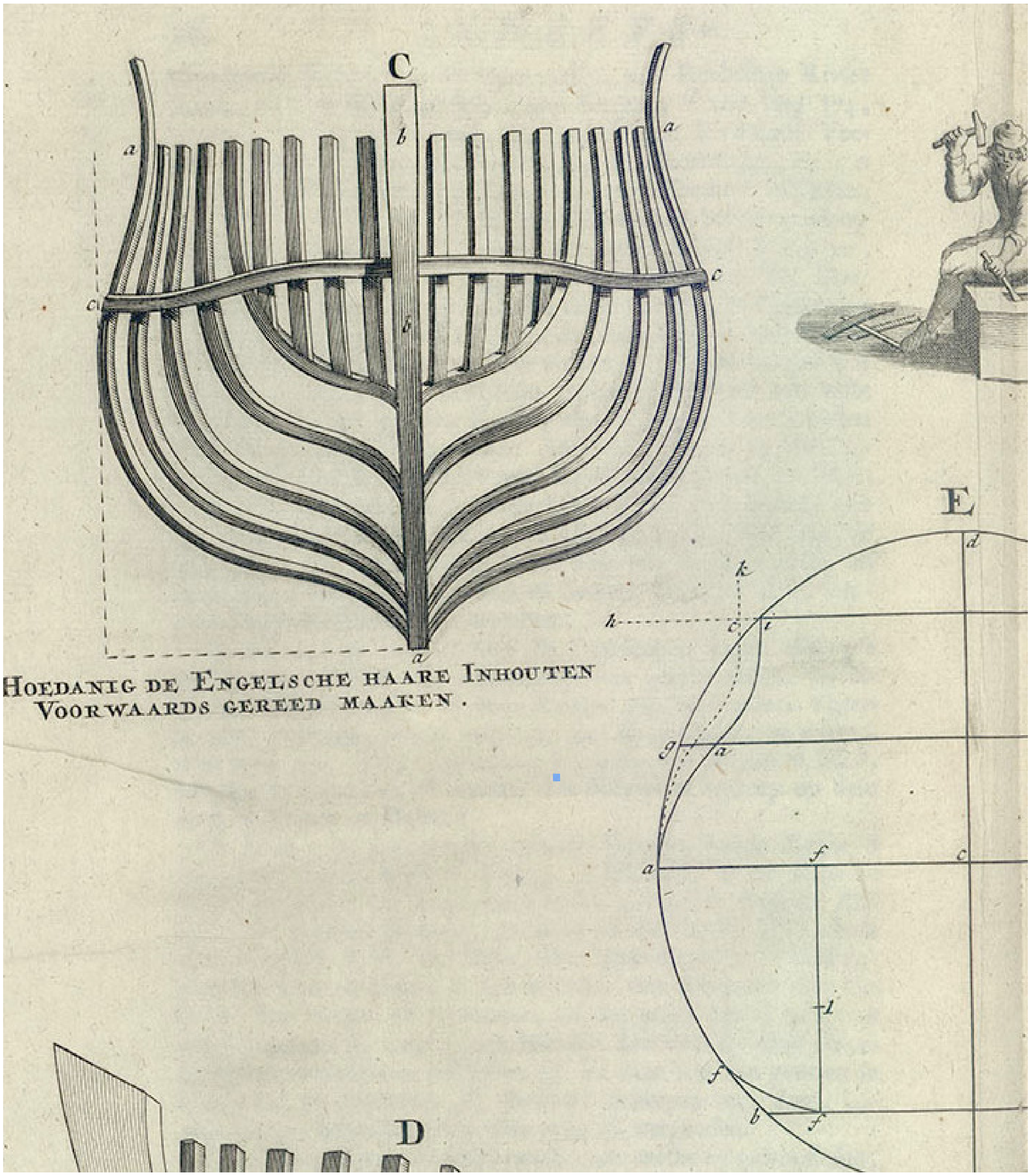


Figure 5.20. Geometric proportions illustrated in a Dutch shipbuilding treatise (Van Ijk 1691).

However, illustrations or visual representations of vessels are not limited to the ones found in shipbuilding treatises. Computer models can be created that depict the components of the vessels. Figures 5.21 to 5.28 show a sequence in the construction of a ship. At each step it is possible to see the transition as new components are added. These figures are the result of the reconstruction of a Portuguese India Nau by [Hazlett 2007]. This hypothetical reconstruction is based on written resources, primarily *Livro Nautico*, complemented with archaeological data.

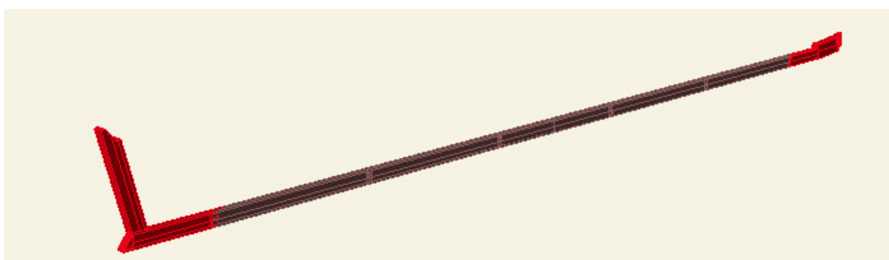


Figure 5.21. Construction sequence (keel). Courtesy Alex Hazlett, CMAC.

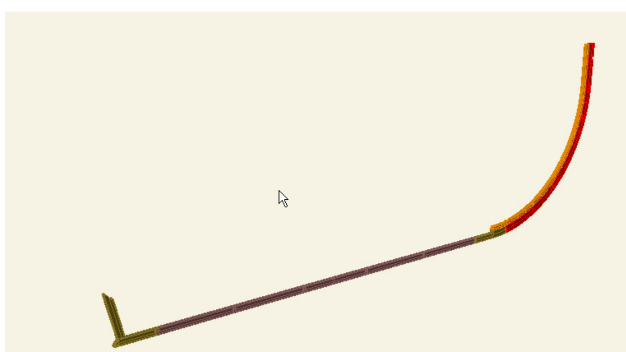


Figure 5.22. Construction sequence (roda). Courtesy Alex Hazlett, CMAC.

Figure 5.23. Construction sequence (codaste). Courtesy Alex Hazlett, CMAC.

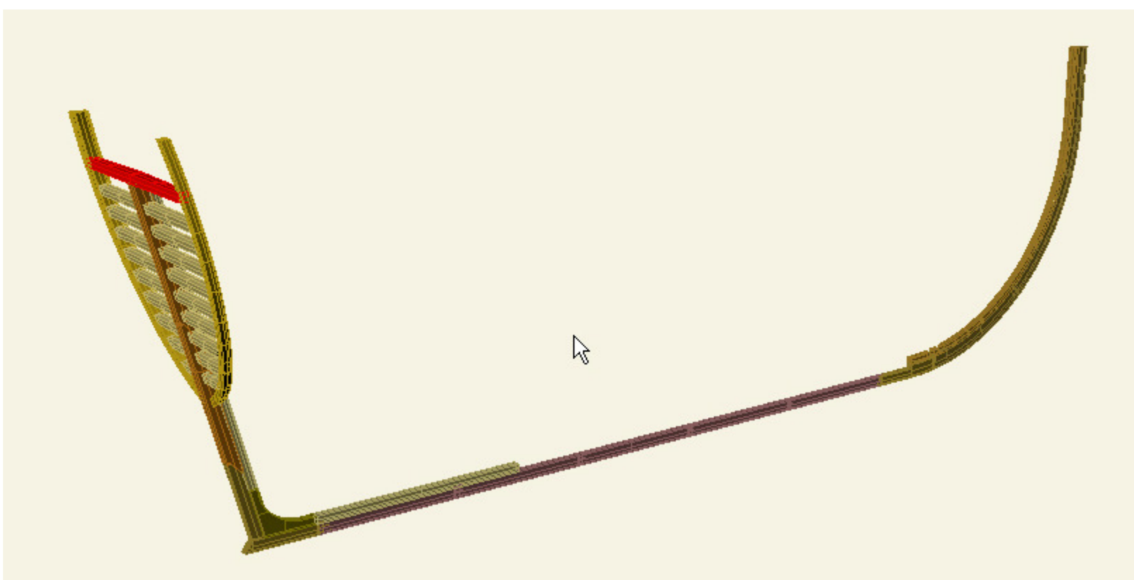


Figure 5.24. Construction sequence (gio). Courtesy Alex Hazlett, CMAC.

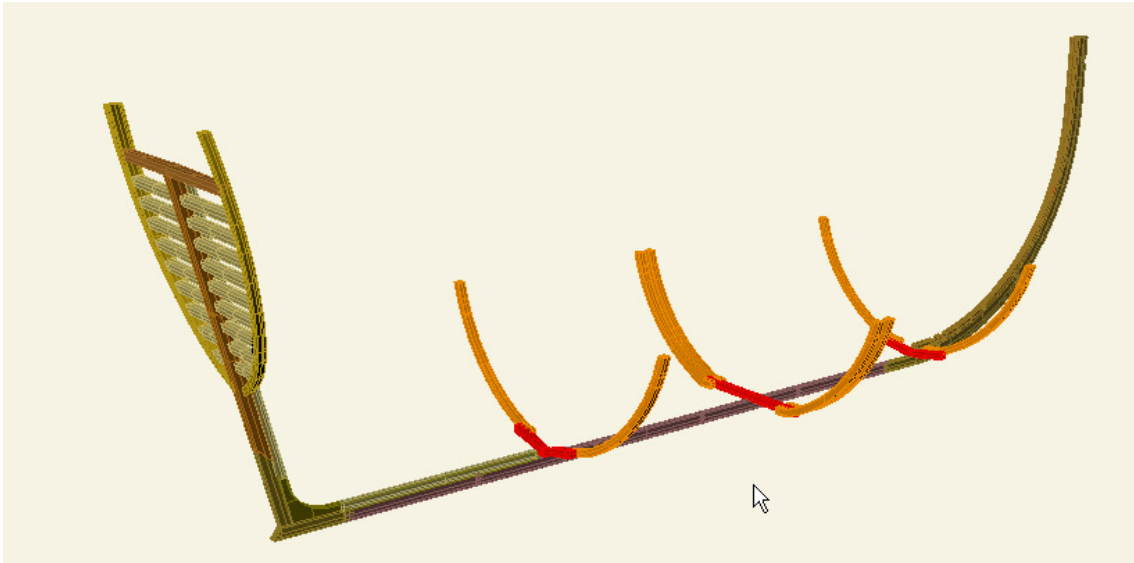


Figure 5.25. Construction sequence (master frame and almogama). Courtesy Alex Hazlett, CMAC.

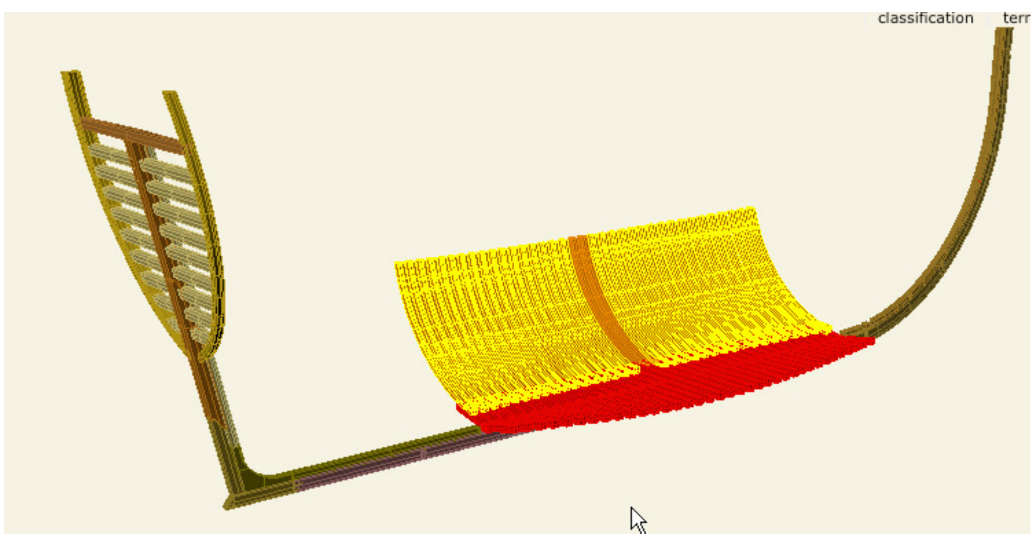


Figure 5.26. Construction sequence (frames). Courtesy Alex Hazlett, CMAC.

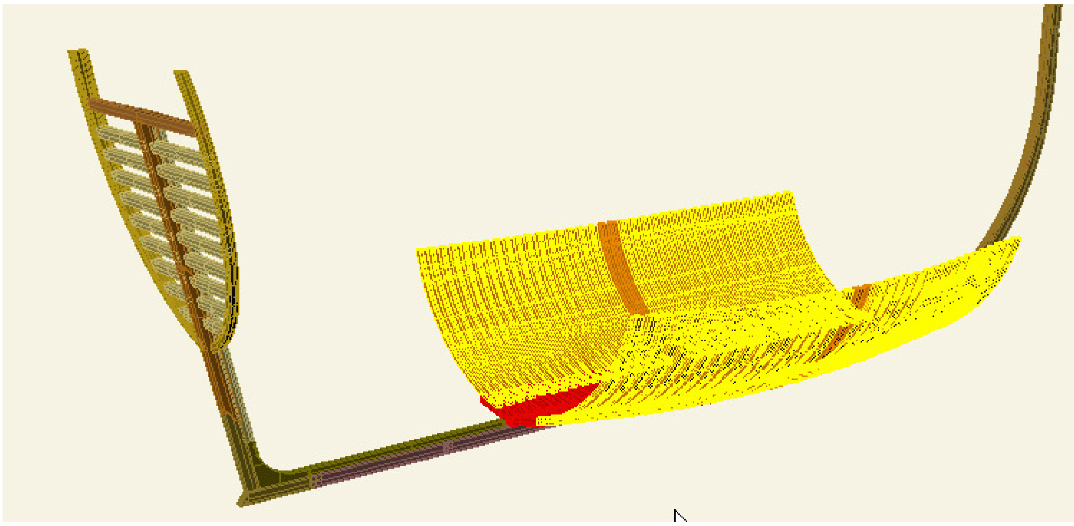


Figure 5.27. Construction sequence (frames 2). Courtesy Alex Hazlett, CMAC.

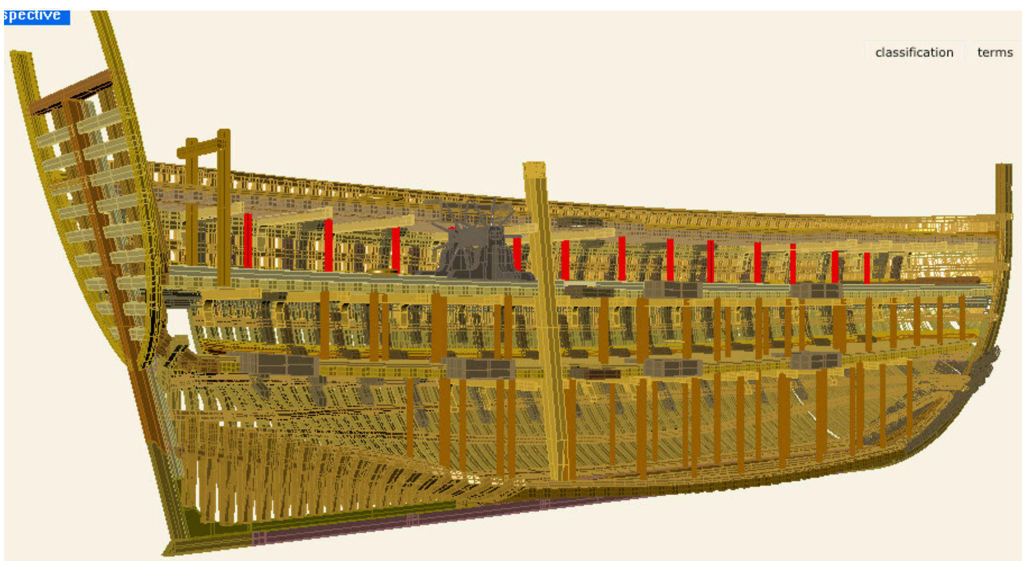


Figure 5.28. A lateral view of a ship in Rhinoceros. Courtesy Alex Hazlett, CMAC.

6. SYSTEM ARCHITECTURE AND *ontoShipDS*

6.1 System Architecture

The architecture I propose is a combination of various well known techniques and tools in digital libraries. Figure 6.1 shows the components of the architecture. Documents (D_1, D_2, … D_k) are encoded in XML and segmented by page. Each document is associated to a language, which allows the use of the proper stemmer in the creation of a full text index. The full text index tool used is Lucene [2009]. Each word in the glossary is stemmed and searched in the full text index. The result is a set of documents containing that word. Associating terms in the text of each document with entries in the glossary requires a parser and a tagger. The parser reads lines of text from the original XML transcriptions. An XSLT template removes existing tags, leaving only words. The tagger searches each (stemmed) word in the glossary. If a match is found, the word is enclosed in a tag with the corresponding identification key from the glossary. Words without match in the glossary are left unchanged. The new line is saved in the XML file.

Matching words is carried out not only by searching for a word in a particular language, but instead by expanding the query using *spellings and synonyms*. Therefore, the original term *(id, T, O)* is transformed into *(id, T', O')*, where *id* is the term key, *T'* is the stemmed term, and *O'* are all the stemmed synonyms and spellings. In addition, tagging can be enhanced further by using the definitions from *ontoShipDS*, since the ontology relates words and expresses relationships.

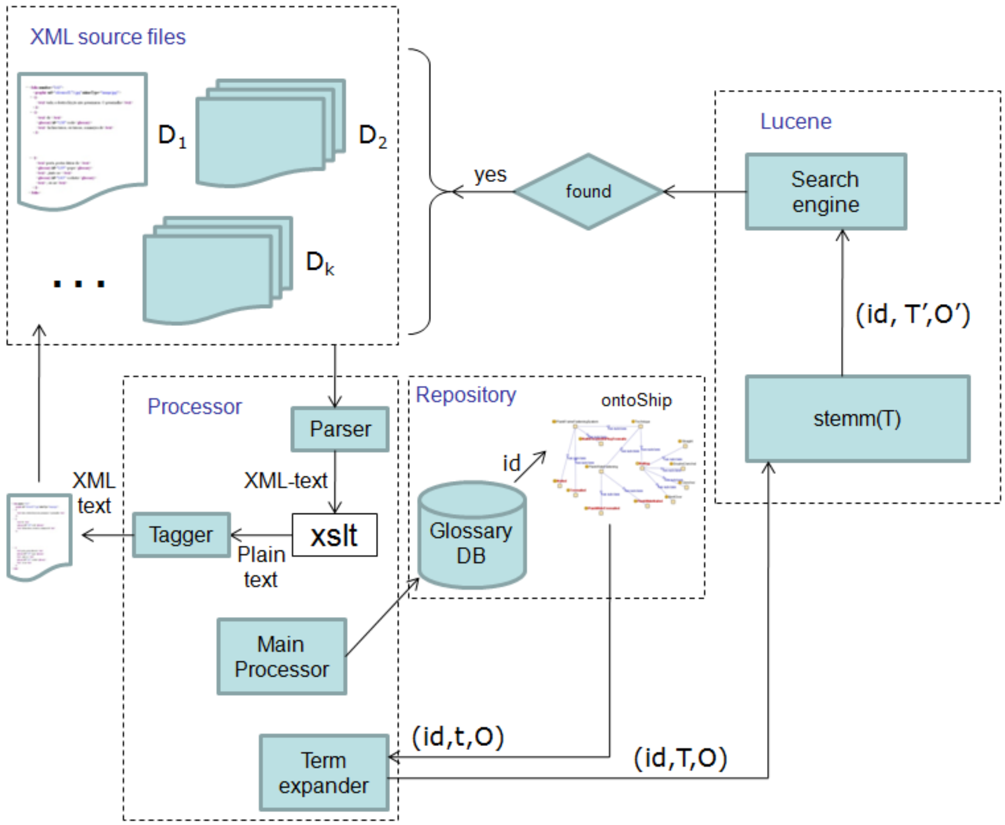


Figure 6.1. A Diagram depicting the system architecture of the framework proposed in this dissertation.

Another component of the architecture is the glossary of nautical terms *glosShip*. This component is based on a relational database (Figure 6.2). In this model, a master table stores unique identifiers for each term. In an alternate table all translations of that term along with its roles (synonyms and spellings) are stored. This includes definitions in each language. Similarly, there are two additional tables for taxonomy and categories. These tables list all the categories and taxa available. Additional tables contain the references between terms and taxonomy, and terms and categories.

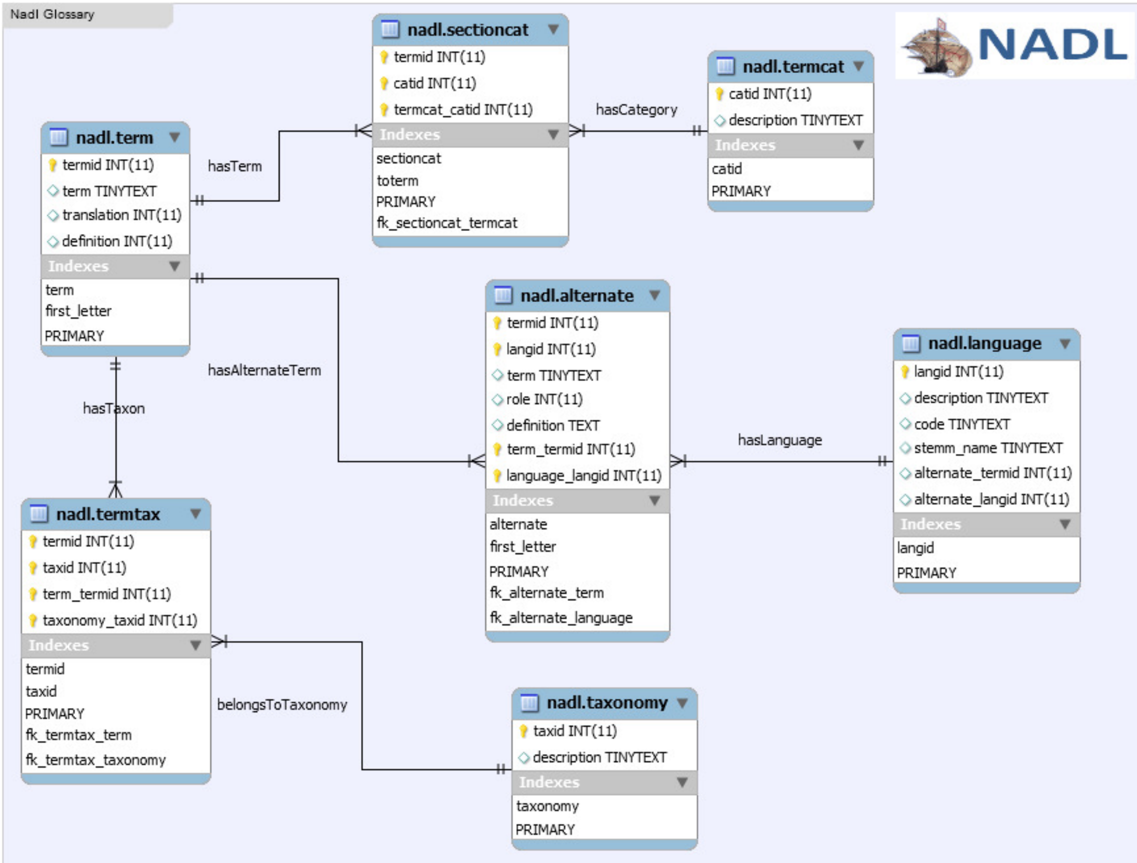


Figure 6.2. An entity-relation diagram depicting the tables and relationships in the implementation of *glosShip*.

The glossary editor is a web-based interface; communication with the server is handled via AJAX. Web-based pages are generated using Java Servlets. Similarly, for the treatises reader interface, the web-based pages are Java Servlets-based. AJAX is also used to request and send information from and to the server. The decision of using AJAX for handling communication with the server was based on the fact that it is a lightweight technology and offers a mechanism for processing XML responses from the server.

Information about images is also stored in a database. A reference table contains keywords (tags) for each image, which allows full text indexing and searching. This option is not limited to photographs, drawings, and illustrations. Images of texts from treatises for which no transcriptions are available can be also tagged. Since tagging is based on terms from the glossary, it offers a controlled vocabulary, which in turn enables location of sections from the treatises based on those terms. Given the degree of specialization and the technical domain of the treatises, this approach seems to offer a viable solution (if not perfect) for documents lacking full textual transcription.

An important component of the architecture is an ontology for describing wooden ships called *ontoShipDS*. This ontology is contained in an RDF-OWL formatted file. It was created using Protégé [Protégé 2000], an open source ontology editor. Navigation through the ontology is possible with a web-based ontology browser. This browser is open source software provided in collaboration between the Universities of Manchester and Stanford [Ontology Browser 2009].

6.2 Introduction to *ontoShipDS*

The word ontology has numerous definitions in different disciplines. In Philosophy for example, [Coffey 1914] offers the following definition:

> Ontology is concerned with the objects of knowledge, with reality considered in the widest, deepest, and most fundamental aspects under which it is conceived by the human mind: …with the modes of its concrete existence and behavior, the supreme categories of reality as they are called: substance, individual, nature, and personality; quantity, space, and time, quality and relation, causality and purpose … (p. 23).

The previous definition helps to frame the discussion of this section. Adding the notion of a community sharing knowledge in a given domain, [Studer et al. 1998] define ontology as: “a formal, explicit specification of a shared conceptualization.”

Based on the experience of using various ontology-editing environments, [Noy and McGuiness 2001] suggest that:

> …an ontology is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions)). An ontology together with a set of individual instances of classes constitutes a knowledge base. (p. 3).

In Information Science, [Guarino and Giaretta 1995] define Ontology as “a logical theory which gives an explicit, partial account of a conceptualization.” [Smith 2003] adds the notion of *canonical syntax* when referring to ontology for knowledge representation. In an attempt to reach a unified definition of ontology for information systems, [Zuniga 2001] contrasts two approaches and shows the differences between them and a philosophical definition.

Situated in a technical domain (mechatronics), [Tudorache 2006] offers a good rationale for the application of ontologies within Engineering Design:

> Some of the benefits of using ontologies are that they define a common understanding of a domain, so that they can be used to support inter-human and inter-organizational communication; and being machine-processable, they support the semantic interoperation of different software systems. (p. 67).

Although ontologies have been created and applied to a wide range of disciplines—each one with diverse goals—what seems to be shared among them can be summarized as follows: a) allow the conceptualization of a domain, b) facilitate software interoperation, and c) improve communication in a community. In this dissertation, *ontoShipDS* accomplishes these goals by first, providing a conceptualization for wooden ships. Second, it establishes the formalization of ship components and their properties; this in turn enables software interoperation. Finally, it creates a framework for documenting and describing archaeological findings for enabling scholarly communication.

The next section begins by describing various examples of ontologies. The second part covers the rationale in designing *ontoShipDS*. The third section describes the main components of *ontoShipDS*. The section concludes with some remarks pertaining to the limitations of the proposed ontology.

6.3 Examples of Ontologies

From Artificial Intelligence to the WorldWide Web, ontologies have become widely used. The Semantic Web is an attempt to add a semantic layer to a document, making its content comprehensible to a computer. Given the growth in the use of ontologies, various ontology-editing tools are now available that facilitate their creation and maintenance. Examples of such tools are: Protégé-2000 [Protégé 2000], Ontolingua [Ontolingua 1997] and Chimaera [Chimaera 2000].

Applications in various domains also abound. A government budgetary ontology along with a methodology for its creation is discussed by [Brusa et al. 2006]. The authors emphasize the importance of defining a robust Domain Conceptual Model. In the legal domain, [Lame and Despres 2005] propose a method for the creation and maintenance of an ontology of the French legal system. Their experiments are interesting since the creation of new laws requires revisions and updates to the existing ontology. This is applicable in nautical archaeology since new archaeological findings can challenge existing knowledge.

OntoLeavn is a text-mining tool whose purpose is to enhance the process of ontology creation [Missikoff et al. 2003]. [Schreiber et al. 1995] discuss an application-driven approach (European ESPRIT project KACTUS) in the creation of tools and methods to be used for the reuse of knowledge during the life-cycle of technical systems. Among the areas in which their approach has been applied include preliminary ship design, oil-production processes, and electrical networks.

In Bioinformatics, [Kawazoe et al. 2008] describe an event ontology for detecting infectious disease-related events reported in textual sources in a multilingual environment. Their solution aims at tackling the difficulty in applying a medical ontology—created by professionals using a highly specialized technical language—to news sources often written in a more colloquial and less technical style. This problem resembles the challenges faced by nautical archaeologists in mapping new information such as characteristics of unknown ship components and construction techniques not previously described in the existing ontology.

Soo et al. [2002] improve image retrieval with the use of a domain ontology and a thesaurus. Unlike traditional keyword-only search, their framework allows semantic queries. First, a parser is used for automatic semantic annotation of images in RDF. During the search process, user's queries in natural language are converted into RDF, this transformed query is then used to match entries in the collection. Their evaluation demonstrates improvements of ontology-based queries over purely keyword-based ones.

6.4 *ontoShipDS*: Describing Wooden Ships

In the creation of *ontoShipDS*, I followed two important principles: "considering the reuse of existing ontologies." [Noy and McGuiness 2001] and "avoiding reinvention" [Soergel 1999]. I started by looking for existing ontologies about ships. The goal was a simple one: to find an existing ontology that could, at least, be used as a starting point. To my surprise I found two: the first one was created by the US Navy for describing basic information about ships [US Navy Ships Ontology 2002]. The second provides an extensive classification for ships and vessels based on tonnage, dimensions, and speed [Commercial Ship Ontology 2005].

However, both ontologies—instead of describing the parts of the ships and their relationships—focus on the characteristics of vessels as a whole. The closest I came to the use of an ontology in shipbuilding, was a proposal for an agent-based architecture in preliminary ship design [Lee and Lee 2000]. The authors neither define nor create a ship ontology: "we have not defined the ontology formally, as it is beyond our scope, and because our prototype agent system is not a large-scale system." However, they seem to acknowledge the importance of an ontology when stating that "The development of a common ontology is essential for a large-scale agent system."

The difficulty in finding an existing ontology of wooden ships can also be attributed to three major reasons. First, vessels are extremely complex machines. Second, modern ships tend to be made of iron and steel instead of wood. Third, the knowledge about ships is quite complex, hence requiring a group of experts in its definition, and considerable amount of time in its formalization.

Nonetheless, what seemed to be a problem in the creation of *ontoShipDS*, shifted my attention in a different yet relevant direction. Instead of attempting to find existing ship ontologies, I began to look at ontologies for describing machines and composite objects. This is based on the fact that ships can be considered machines in that they are made out of numerous components (each one with particular properties and functionalities), which interact and are related to each other. Further, these components have to satisfy certain constraints.

A good example of the use of ontologies for describing machines is the work of [Tudorache 2006] in the creation of an ontology for improving the design of complex engineering devices. In short, the scope in the creation of *ontoShipDS* (an ontology of wooden ships) lies at the intersection of three disciplines: nautical archaeology, computer science, and information science. Nautical archaeology provides the domain knowledge. Computer Science and Information Science offer the theoretical background, methodology, and tools.

In this context, Tudorache's doctoral dissertation focuses on the use of ontologies in an engineering process, the one of building a complex mechatronic device: a car's automatic transmission. Two of the major complexities in her study include the interaction of multidisciplinary teams of experts (mechanical, hydraulic, electrical, and software), and the variety of abstractions or models generated for the device under development (geometrical, functional, multi-body, hydraulic, and software). This fact is illustrated when the author asserts that:

> However, checking the consistency of a design model and between different design models is a very challenging task. One of the reasons is that the engineering tools employed by the development branches operate on models that have different conceptualization of a product according to their own viewpoint. For instance, a computer-aided design (CAD) tool will model the geometrical characteristics of a product, while a simulation tool will model the behavior of the product. Although the two design models represent different viewpoints on the product, they must be consistent with each other, if a common implementation (i.e., the product) of the two design models should be realized. (p. iii).

A fundamental difference between the approach presented in my dissertation and Tudorache's is that the latter focuses on the design of a new physical object, while in Nautical Archaeology the object (ship) has already been built. It sunk, and now has to be rebuilt from incomplete and damaged archaeological evidence. In spite of the differences, both approaches work with composite physical objects, and various abstractions, hence making them very alike.

The ontology I propose is based on the RDF/OWL model. Although ontologies can be described in different ways, RDF/OWL was chosen mainly because it is a W3C standard for the semantic web and the possibility of being expressed in XML. A similar argument can be found in an ontology for modeling information in historical documents [Ciula et al. 2008].

6.5 The Methodology

The first consideration in the creation of *ontoShipDS* was adopting a methodology. Although a "correct" methodology does not exist [Noy and McGuiness 2001], a common practice seems to indicate an iterative approach; one in which an initial version is further revised and refined. However, a few "dogmatic" rules in ontology design are proposed by Noy and McGuiness:

> 1) There is no one correct way to model a domain— there are always viable alternatives. The best solution almost always depends on the application that you have in mind and the extensions that you anticipate.
> 2) Ontology development is necessarily an iterative process.
> 3) Concepts in the ontology should be close to objects (physical or logical) and relationships in your domain of interest. These are most likely to be nouns (objects) or verbs (relationships) in sentences that describe your domain. (p. 4).

Because I am not an expert in Nautical Archaeology and the domain is complex and large, the first obstacle in designing *ontoShipDS* was "to find" the knowledge. Preliminary conversations with Dr. Filipe Castro indicated three sources of knowledge: textual, visual, and archaeological. Textual knowledge was found in a collection of shipbuilding treatises describing the construction of ships. Visual knowledge was found in a collection of photographs depicting ship timbers and components documenting hundreds of excavations along with illustrations in shipbuilding treatises. Archaeological knowledge on the other hand, came from a compilation of data and information generated by archaeological excavations of hundreds of ships and stored in a relational database.

Textual and visual knowledge are discussed in Sections 4 and 5 respectively. In the following paragraphs I will describe the work of Mr. Steffy—a renowned nautical archaeologist and researcher at CMAC—in the compilation of information and creation of the database previously mentioned. The database was an attempt to enhance the storage and dissemination of information pertaining to hundreds of excavated and surveyed shipwrecks spanning decades of work. In his own words, Steffy explains the reason in the creation of the database as follows [Steffy 2003]:

> This compilation of hull construction information is intended to address the hull remains of wrecks found in any geographical area and dating to any period. It is largely based on the reports of shipwreck studies and consists of a series of relational database files… The basic purpose of the project is to assist archaeologists and archivists in evaluating and enhancing their contributions to the history of wooden ship and boat building technology. Consequently, the contents address only structural wooden displacement vessels—i.e., assemblies of wooden hull components. (p. 1).

Writing about the information and data generated from archaeological excavated shipwrecks and the need for its preservation and dissemination he adds [Steffy 1994]:

> …Piles of rotted timbers and broken artifacts constitute a wealth of information, yet much of that knowledge will remain unrecognized unless one develops a proper method of access to it. In the case of shipwrecks, though, access is the mastery of a discipline, which is essentially the means of access to the wealth of information stored in the remains of ships and boats and the orderly dissemination of the knowledge derived from them… (p. 189).

Working in a field that yields large amounts of data and information, Steffy anticipated how this wealth could profit from the creation of and storage in a database. In other words, he understood the role of adopting computing technology in its preservation and dissemination. This notion led him to create a relational database in Microsoft Access for storing information generated by numerous archaeological excavations. As expected, his original entity-relation model is not a normalized one, hence contains inconsistencies. Furthermore, additional information such as notes and publications was stored in plain Microsoft Word documents.

I argue that Steffy probably did not realize that his database design, although flawed from the database theory standpoint (he did not have a background in computer science), offers a unique depiction of the knowledge about the properties and relationships of ship components. This illustrates how an attempt to automate the storage of archaeological-generated data, turned—probably by accident—primarily into a mechanism for the preservation of knowledge that had otherwise been almost impossible to achieve. Steffy's pioneering work provided the building blocks for the creation of an ontological representation and conceptualization of wooden ships.

In contrast, computer science provides a methodology for formalizing the conceptualization of wooden vessels. The creation of *ontoShipDS* started from a preliminary analysis of the structure of the database. To facilitate the editing process I used Protégé [Protégé 2000]. Dr. Filipe Castro was the domain expert. Throughout various sessions, unknown concepts were clarified. Castro's explanations were instrumental in removing redundancies and inconsistencies, as well in ensuring proper classification. Entries in the model were categorized either as *concepts* or *objects*. The modified version was reviewed and critiqued by Castro. What emerged from this process was a preliminary conceptualization of wooden ships. The contents of the database on the other hand, both recorded hundreds of ships and provided priceless instances of real components and features from numerous vessels, hence augmenting and offering an initial validation of the conceptualization.

In order to define the scope of an ontology, it is important to formulate a list of questions that the ontology should answer [Grüninger and Fox 1995]. These questions are known as "competency questions." As expected, in the case of *ontoShipDS* it was difficult to know all possible questions a priori. This is due in part to the size and complexity of the knowledge about wooden ships. Nevertheless, the following paragraph lists a few questions that provide a general idea of what the ontology should aim at answering.

What kind of planking technique was used in a ship given the fastening objects found in an excavation? What were the frame dimensions in a ship based on a fastening technique? What planking pattern do holes and cuts in a timber suggest? What is the best joining method for a particular planking? Is it possible to see all assembling techniques for certain timbers? What can be learned from the wood used in building a vessel? What are the best types of woods for the various parts of a vessel? What do dimensions of a recovered piece suggest in terms of the size of another timber that was not present?

As stated earlier, it would be impossible to cover all possible competency questions for this domain. However, a good way to summarize what the ontology should be able to answer is as follows: to depict as many as possible the components of a wooden ship; to describe their properties, how they are assembled, and what objects are used in their assembly; and to present instances of ships that document the techniques previously listed.

The preliminary question regarding the creation of an ontology as part of this dissertation can be stated as follows: If ontologies have been extensively studied and used in computer science and information science, what is the justification for creating *ontoShipDS*? Although arguing that *ontoShipDS* can facilitate the work of nautical archaeologists and enhance the dissemination of their findings would suffice, it falls short from its full potential. A more compelling explanation is the fact that the majority of ontologies describe an existing knowledge, even if new facts and constrains are later added, or existing ones need to be modified or removed. *ontoShipDS* in contrast, expresses a changing knowledge, since it describes components of unknown, damaged, and incomplete vessels.

This changing knowledge is at the heart of archaeology since new discoveries can require the modification of existing concepts that were previously known, as new archaeological evidence emerges. For practical purposes, *ontoShipDS* can also help archaeologists to investigate hypothesis about properties of a particular component that are not explicit. For instance, an archaeologist working on a given timber with limited or no knowledge about its properties and relationships limits the scope of his or her investigation to what is explicitly known, probably leading to missing important research avenues.

Although for an expert this might be not a frequent situation, as it is assumed that he or she would know at least most—if not all—of these properties and relationships. In nautical archaeology, with new discoveries being constantly made on each excavation, even experts are subject to missing important new facts. For the general public and non-experts, on the other hand, *ontoShipDS* can be used as a discovery engine, by presenting implicit facts and properties.

ontoShipDS can also be used as a tool for comparing or complementing the representation about an object and its parts expressed in different media; in the present dissertation, textual and visual representations of ship components. Nautical archaeology again offers an interesting case, because a textual description of a ship component in a shipbuilding treatise might fail to show properties that are explicit in a photograph or illustration. Conversely, a property or feature in a photograph or illustration might be difficult to understand without a written explanation. The reason I argue that *ontoShipDS* can help in bridging this gap is because both visual and textual abstractions correspond to the same physical object, thus facilitating and assisting archaeologists in the reconstruction and understanding of wooden ships.

The motivation for creating an ontology of ships also coincided with my work in the design of a computer-based infrastructure for shipbuilding treatises [Monroy et al. 2007b], along with a prototype for storing and tagging images and illustrations of ships. Examining written descriptions of vessels in the treatises revealed interesting characteristics about ships that can only be conveyed in this media. Conversely, illustrations from the treatises, drawings, and photographs from underwater excavations offer a different abstraction of a vessel, revealing other properties that a written description cannot accomplish.

6.6 *ontoShipD*S

The first step in the creation of *ontoShipDS* was the identification of the main categories. With an original representation captured in the form of a relational database, tables provided a natural categorization scheme. This scheme however, was not free of inconsistencies, mainly because concepts, objects, properties, and techniques were mixed within the same tables. Despite these imperfections, the relational database encapsulated the core concepts, such as: main structural components, auxiliary components, and various joinery systems, types of wood, diverse objects, techniques, and measurements.

As stated earlier, building an ontology is an iterative process, one in which domain experts are essential in providing the knowledge and validating that the model is sound. In the following sections I explain the transition from the original database conceptualization into an ontological formalization.

6.6.1 Original Conceptualization

Because the contents of a database can be represented as an entity-relation model, it offers a natural environment for describing components and properties in a domain. It is not a surprise that Steffy found in a database (Microsoft Access) the framework in which the knowledge about wooden ships could be initially represented.

Despite the benefits databases offer for such representations, they have various disadvantages. For example, they are heavyweight, complex relationships are difficult to be represented, sharing and authoring is not an easy task, and reformulating relationships can be difficult. In short, they are not designed for a rich knowledge representation, although they can serve as a foundation for it. However, it is fair to say that great progress has been made in mapping databases to ontologies [Trinkunas and Vasilecas 2007; Wiesman et al. 2002].

The original database entity-relation model can be seen in Figure 6.3. Note the difficulty in grasping the relationships among components as well as the use of abbreviations to describe attributes. Although this model was accompanied with a narrative describing in more detail the contents of each table, and some relationships, it is still difficult to fully understand the knowledge it encapsulates. However, both the ER-model and the narrative served as starting point for *ontoShipDS*. The contents of the main tables are described in the following paragraphs.

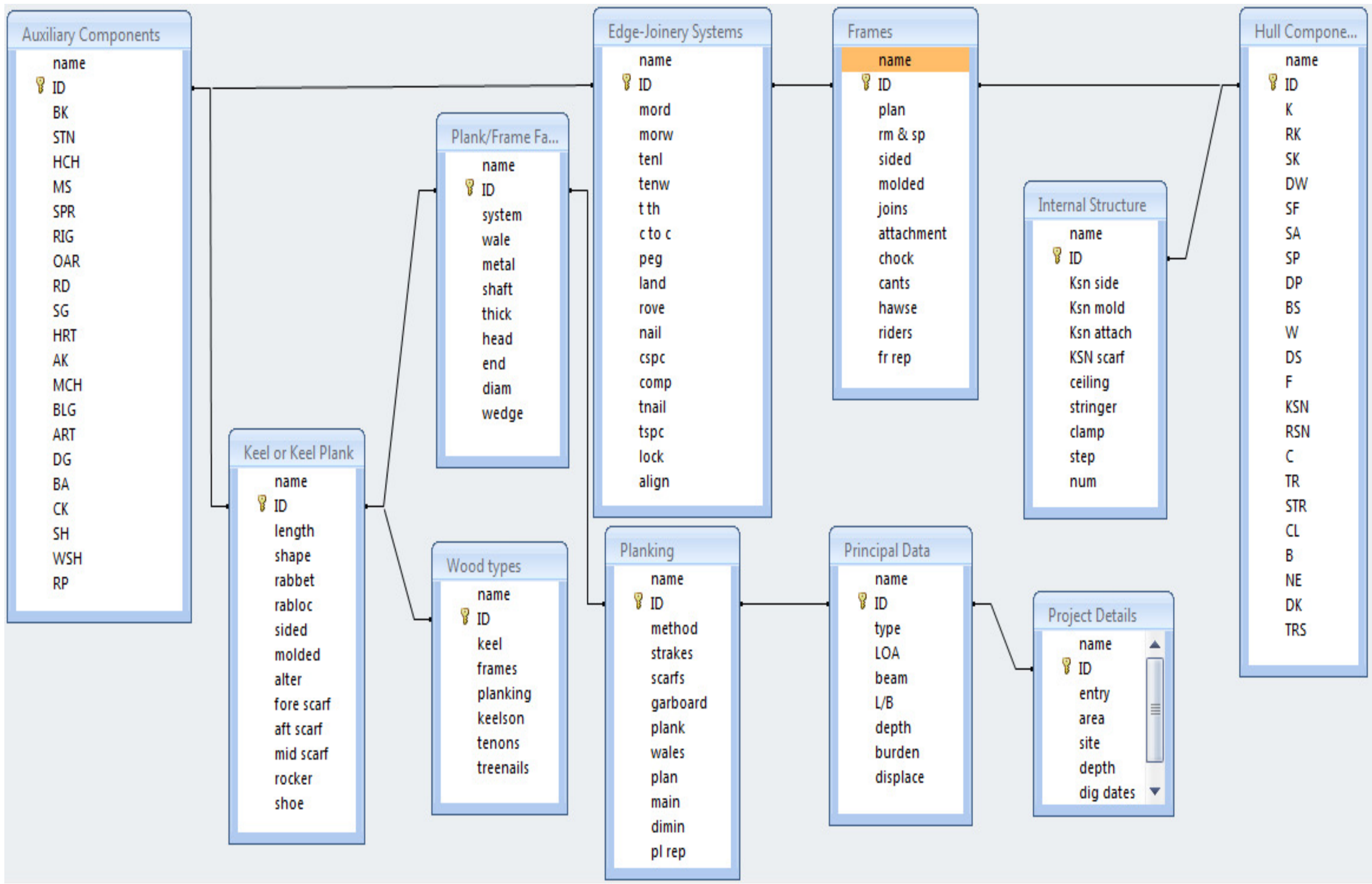


Figure 6.3. Original entity-relation model of Dick Steffy’s database.

Principal Data: This table contains a list of excavated or surveyed ships, with general description and dimensions of the ship such as type of vessel, length, breadth, height, and ratios.

Hull Components: "a list of surviving hull timber fragments found on the wreck." This table covers structural components of the hull. An initial inspection of this table showed that five out of twenty-three of the original attributes were not really components but rather concepts related to the hull. After further analysis and discussions with nautical archaeologists, it became clear that it was more appropriate to call them techniques instead of concepts. In addition, two other items had their own separate tables, suggesting implicit relationships.

Auxiliary Components: "list of surviving secondary components or component fragments found on the wreck." Unlike hull components, elements described in this table do not play a structural role. Rather as the name suggests, they can be seen as complementary. Similar to the previous table, this one included various attributes that were more suitable to be categorized as techniques. Other attributes provided elements belonging to various categories.

A question that naturally emerges from the contents of these two tables in the context of ship reconstruction is what the relationship among components is. They are related in that characteristics of one component can lead us to hypothesize on something that was not found. For instance dimensions, cuts, and shapes of a mast step can indicate the size of a mast for which no archaeological evidence was found. This in turn, can help in understanding the dimensions of the ship, for which probably just a fraction of the hull was recovered.

Wood Types: "generalized identification of wood types used in the fabrication and assembly of major structural components." Originally, this table listed the most common wood types used in the major structural components. In *ontoShipDS*, this was expressed as a class **Wood** and a property **hasWood**. This approach allows adding the property to any component of a vessel. For instance *hasWood*(Keel, Oak) states that a Keel is made out of Oak.

Keel or Keel Plank: this table lists characteristics of the keel and keel planks such as dimensions and shapes. It also describes materials and objects used in connecting keel and keel planks. Some items in this table were reclassified as techniques and properties respectively.

Frames: the structure of this table was very similar to Keel and Keel Plank. Some of the information stored in this table includes properties, characteristics and dimensions. It also provides various techniques for attaching frames to the keel.

Planking: this table describes planking methods, joining techniques, dimensions and other properties.

Plank/Frame Fastening Systems: this table contains information about the methods used in fastening planks with frames. It also describes the objects used in fastening along with their properties and dimensions.

Edge-Joinery Systems: this table describes techniques and objects used in joining planks to planks. It also includes properties and dimensions.

Internal Structures: In contrast to outside structural elements, this table contains characteristics and information about structural timbers located inside the ship (including dimensions).

6.6.2 Transforming a Database into an Ontological Representation

The transformation process started with a careful analysis of the structure of the database, classifying concepts that represented similar ideas. This was a fruitful method because it allowed the consolidation of repeated information. In addition, it provided the foundation for the definition of the top level categories in the ontology.

Six high level categories were identified: Component, Concept, Object, Property, Technique, and Ship. **Component** describes any physical object that is a structural part of the ship. **Concept** contains general information about a vessel, such as its type and dimensions. **Object** includes objects that do not play a structural role in the ship, but are necessary to fasten, join, and attach them. **Property** describes the materials ship components are made of; for example, wood and metal. **Technique** lists all techniques used in building the ship. Techniques are divided into categories based on the components of the ship where they are employed; some examples include: Ceiling, Chocking, Nailing, and Planking. **Ship** contains information about (wooden) ships excavated or surveyed by CMAC and their general characteristics.

The original structure of the category Hull Components given in Steffy's original database, illustrates part of this transformation. Attributes such as **single** and **double planking**, **common** and **transverse ceiling** had originally their own entry in the table. Two things were immediately evident in this case, first that they belonged to two different categories **planking** and **ceiling**. Therefore, they were converted from a non-categorized representation into a categorized one. However, further discussions with domain experts suggested another conceptual ambiguity. If this table described components of a ship, the terms discussed here—ceiling and planking—were not components as such, but rather techniques. In fact, this case was found in a number of tables, hinting the creation of a new category: **Technique**. This new category was used to encompass any type of technique.

A second observation indicated that a few attributes in the original table had separated individual tables. Each table, in turn, included its own list of properties. This was the case of **keel**, **plank**, and **frame**. What emerged at this point was the need for the creation of a new category that summarized components in general. This finding coupled with the contents of other tables, lead to the creation of **Components**, which in turn was broken down into: Hull Component, Auxiliary Component, and Internal Structure.

Another characteristic shared among objects in numerous tables was the material they were made of; this case, metal and wood. This category was named **Property**. Wood was then divided into various wood species used in the components of a ship. This makes possible to represent the fact that any given component of a ship is made of a species of wood.

As indicated earlier, all components that did not play a structural role on the ship were classified under **Object**. Similar to **Property**, this class enabled the association of objects to various structural components. For example, there are various techniques for fastening a frame to a plank; each technique uses different objects such as nails, treenails, and bolts to name a few.

Figure 6.4 shows the original definition of the table **Plank/Frame Fastening Systems**. As the name suggests, this table contains information about the techniques used in fastening **Planks** and **Frames**. Of the contents in this table, only items 2 and 3 correspond to techniques. Item 4 is a property that applies only to metallic objects (nails). Items 5, 8, and 10 are properties that applied to certain objects described in 2. For example **end** refers to the type of technique used in placing the nails. This technique however does not apply to **treenails**. Finally items 6, 7, and 9 describe measurements of certain objects. Item 6 refers only to **shafts** (item 5), item 7 only to **nails** (included in item 2), and item 9 to **treenails** (included in item 2). What this example illustrates is that a database enables having an initial understanding about a concept. However, it also shows the inconsistencies in describing properties, as well as the difficulties in finding relationships. For example, item 3 states that a **wale** can be fastened using a different method than the one used in fastening frames with planks.

XI. Plank/Frame Fastening Systems

1. ID (vessel ID)
2. system (type of fastening system – [N] for nailed; [T] for treenailed; [N/T] for combinations; [L] for ligatured; [PT] for nails encased in plug treenails; [B] for bolted; [O] for other; [X] for unknown
3. wale (type of wale fastenings, if different than plank fasteners)
4. metal (type of metal)
5. shaft (square, round, or other)
6. thick (shaft thickness or diameter, in cm)
7. head (head diameter or maximum dimension, in cm)
8. end ([CL] if the end of the nail was simply bent over; [DCL] for double-clenched applications; [S] for straight nailing)
9. diam (treenail diameter, in cm)
10. wedge (if a treenail, was it wedged? yes or no)

Figure 6.4. Original definition of the table plank/frame fastening systems.

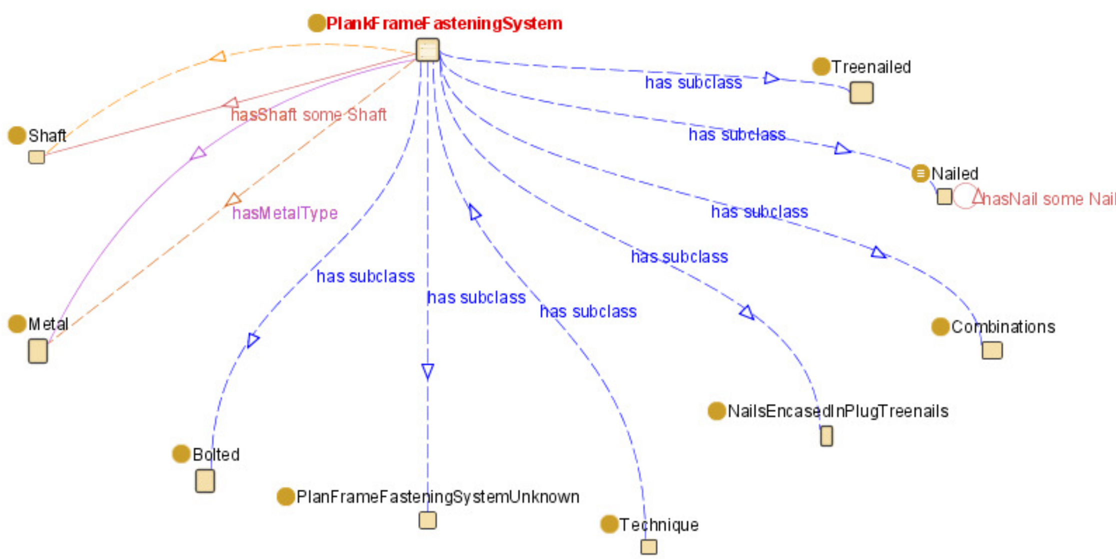


Figure 6.5. A visual representation of plank/frame fastening systems from *ontoShipDS*.

Unlike a database depiction, an ontological representation enables a richer description of concepts. Figure 6.5 shows a graphical representation of **Plank/Frame Fastening System** from *ontoShipDS* in Protege. There are various facts that can be learned from this representation. For instance, the direction of the arrows indicates membership to classes. For instance **Plank/Frame Fastening System** is a subclass of **Technique**. In turn, there are various **Plank/Fastening Systems**: **Treenailed**, **Nailed**, **Combinations**, **NailsEncasedInPlugTreenails**, **Bolted**, and **Unknown**. Also, it can be inferred that some fastening techniques use shafts, and are made of metal.

A more detailed description is shown in Figure 6.6. This shows for example that **Planks** and **Frames** can be **Nailed** together. Further, Nailed as technique requires Nails, and that there are at least five techniques for securing nails (**Bent Over, Clenched, Double Clenched,** and **Straight**). Additionally, Planks and Wales can be fastened using Treenails or Nails. In turn, **Treenails** are made of wood. Finally that **Treenails** can be **Wedged**, and that a **Nail** has a **Head** and an **End**. These two examples help to contrast how *ontoShipDS* offers a richer description of concepts in a domain than the one provided by a relational database.

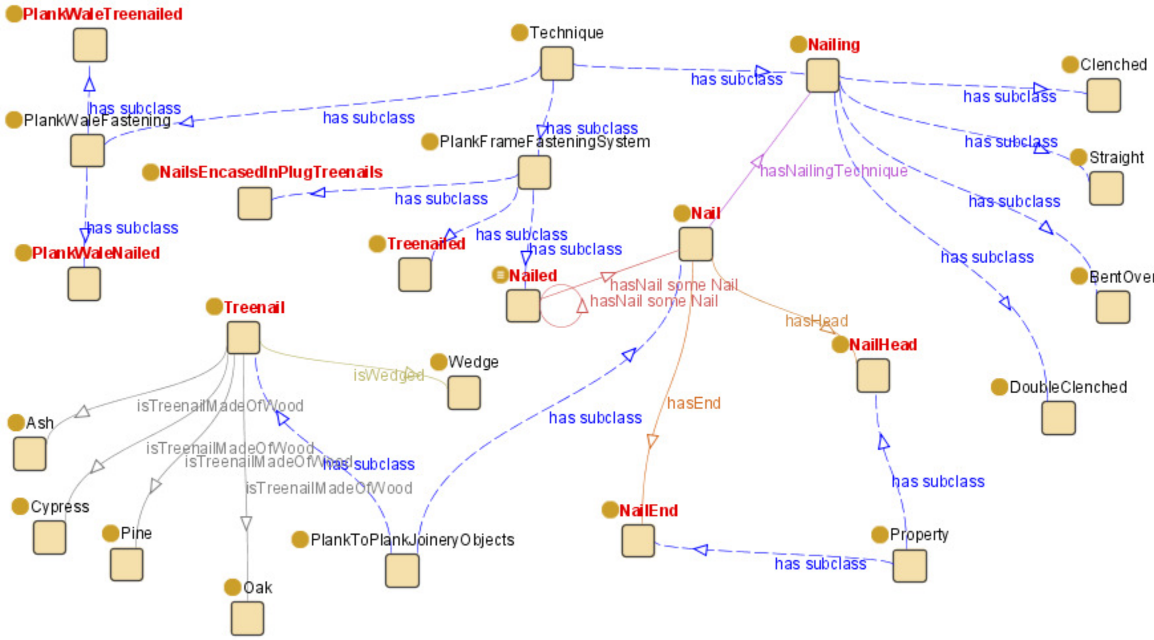


Figure 6.6. An extended visual representation of plank and frames fastening techniques.

Also characteristics and properties of ship components and objects are subject to being described in an ontology. For example, a ship can be described in terms of certain properties such as general dimensions, which include: length, breadth, length/beam ratio, depth, and displacement. In *ontoShipDS*, these properties were categorized under **Concepts**. Although it can be argued that these categories could be considered concepts (and conceptually they are), the adopted categorization suits the needs of this domain.

The final top-level class is **Ship**. This class includes instances of vessels either excavated or surveyed by CMAC. These instances are then associated to the rest of the classes in the ontology. This enables to show concepts about ship constructions from real examples.

Throughout the construction of the ontology there were difficult concepts to classify, for example, Nail, Nailed, and Nailing. Nail was classified under Object. However, Nailed and Nailing proved to be more difficult. Initially, they both seem to represent the same concept. Again, consultations with the domain experts clarified this apparent ambiguity. They both were classified under Techniques. Nailing describes the techniques in which a nail can be inserted and secured. Nailed on the other hand, is a type of Technique for fastening a Plank and a Frame.

In summary, nautical archaeology offers an interesting environment in the use of ontologies for ships. One of the main reasons is the difficulty in the creation of a canonical representation of a wooden vessel. However, this does not imply that an ontology for ancient wooden ships is not achievable; on the contrary, it is advisable.

7. USES OF *ontoShipDS*

Ontologies can be used for various purposes. For instance, one of the most common applications is the formalization and conceptualization of a domain [Gruber 1993]. Therefore, the first application of *ontoShipDS* is in the formalization and conceptualization of ships and their components (limited to wooden vessels). The emphasis is made on how the ontology augments their visual and textual representations. Because rebuilding ancient vessels relies on comparing ship fragments and timbers from underwater excavations with other archaeological evidence, written materials, and technical illustrations, *ontoShipDS* can be used to augment queries enhancing the comparison of written and visual abstractions of ships. In this section I discuss various scenarios in which *ontoShipDS* can be used.

7.1 Formalization and Regularization of Concepts

The first application of *ontoShipDS* is the formalization of the knowledge about wooden ships. Because properties and relationships among components are well described, *ontoShipDS* facilitates the understanding of ship parts. It also enables regularization by providing a classification mechanism that guarantees consistency. Classification ensures that instances of classes are properly identified and categorized according to their roles and properties.

Another advantage of an ontological formalization is the possibility to express properties, constraints, and relationships about objects in a domain. One method for representing this formalization is description logic (DL) [Baader and Nutt 2003]. OWL (the official W3C ontology representation), for example, uses Description Logics. The following paragraphs show the use of DL in the formalization of concepts. Each example includes a photograph, a traditional dictionary-like definition, a DL description, and two visual representations.

Before describing these examples, it is important to explain two concepts used in DL: sufficient and necessary conditions. A necessary condition, states that if an instance belongs to a given class, it has to satisfy that condition. In other words, for an object to belong to a given class it is *necessary* to satisfy that condition.

In contrast, a sufficient condition states that if an individual satisfies that condition, it can be determined to be a member of that class. This is to say, that the condition is *sufficient* to determine that object's membership to a class. However, this does not imply that an object cannot belong to other classes. In fact, objects most often belong to multiple classes.

The first example is **Plank**. A plank can be defined as: "A timber used in the outer lining, or shell of a hull." The following statements provide a formal definition of **Plank**. Each statement is expressed as property(Object, Domain).

isA(Plank , HullComponent); (Definition 7.1)
hasWales(Plank , Wale);
hasStrakes(Plank , Strake);
hasPlankWaleFasteningSystem(Plank , PlankWaleFastening);
hasGarboard(Plank , Garboard);
hasPlankingPattern(Plank , PlankingPattern);
hasPlankingTechnique(Plank , Planking);
hasPlankingPlan(Plank , PlankingPlan);
hasPlankReplacements(Plank , PlankingReplacements);
hasPlankRepairs(Plank , PlankingRepairs);
hasPlankingScarfs(Plank , Scarf);
hasPlankToPlankFasteningTechnique(Plank , PlankToPlankJoinerySystem);

Figure 7.1 depicts a plank from the *Pepper Wreck* excavation along with an embedded definition. Combining the definition with the photograph it is possible to have a general idea of a plank as a component of the hull of a ship. However, nothing else about planks can be learned, except perhaps that they are pieces of wood, and seem fairly strong. Additional information such as to with what other components they can be assembled, or what objects and techniques can be used for their assembly are not known. Also, it would be useful to identify what species of wood were the best for planks, and provide examples of ship planks from various excavations.

Figure 7.1. A plank from the *Pepper Wreck* excavation with an embedded definition. Photograph courtesy of CMAC.

Objects can also be depicted in different contexts. Figure 7.2 shows a drawing of a plank depicting scarfs, nail holes, curvatures, and a scale. This drawing shows important details that are hidden in the photograph. Another example can be found in Figure 7.3, which depicts (in red) planks of the first deck in a 3D model of a Portuguese India Nau. In contrast to Figure 7.1, this alternate depiction offers a context of where in the ship planks are placed. Although some details from the photograph are lost, these two depictions offer additional information about planks. For instance, in what parts of the vessel they can be found, and their estimated proportions in terms of the breadth of the vessel. Despite the additional knowledge introduced with these two visual abstractions, there are still remaining unknown concepts and relationships about planks.

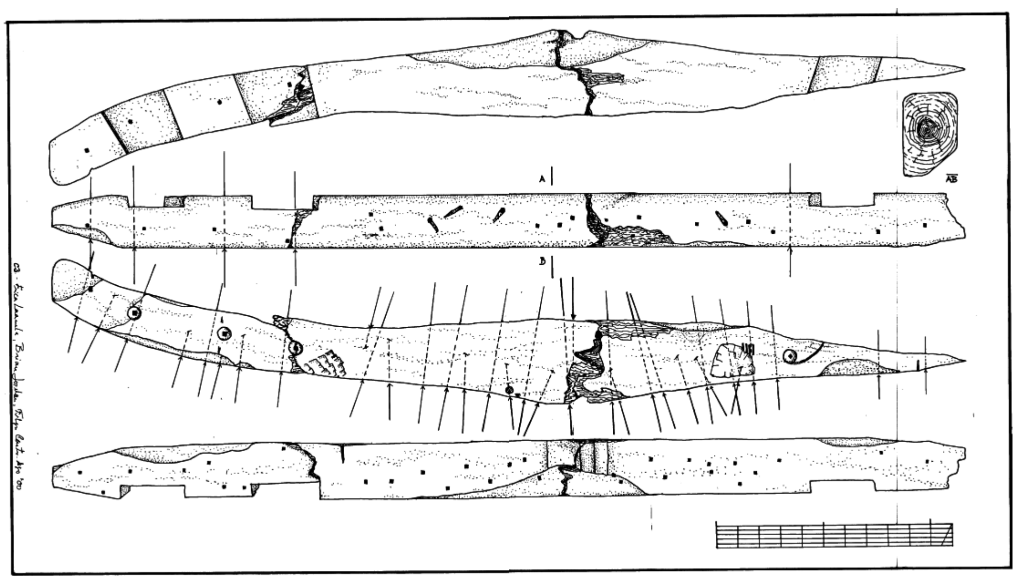


Figure 7.2. Drawing of a plank with details. Courtesy of Dr. Filipe Castro CMAC.

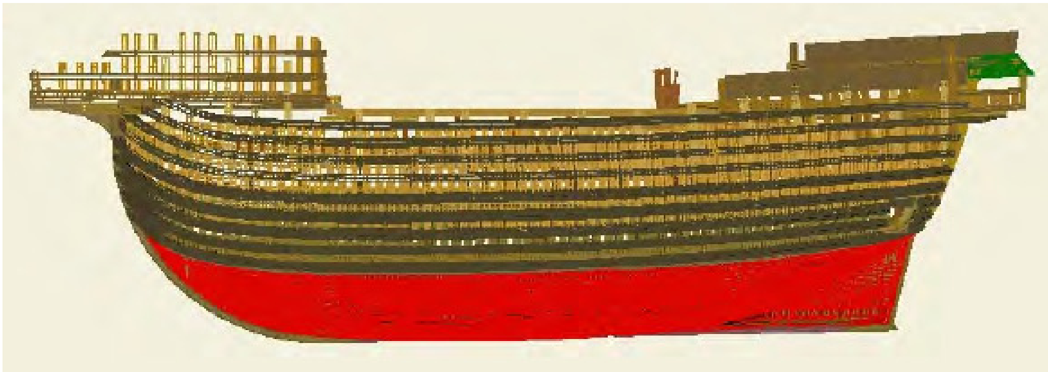


Figure 7.3. The lower planks (in red) of the hull of an India Nau. Model courtesy of Alex Hazlett, CMAC.

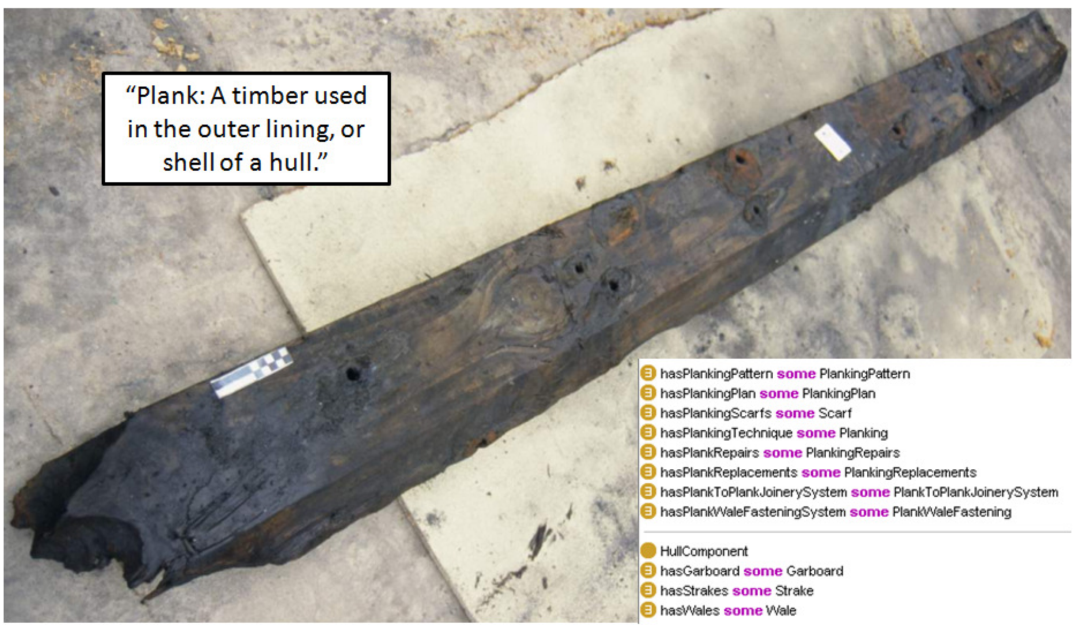


Figure 7.4. The same plank depicted in Figure 7.1 along with its definition, and a list of conditions in description logic.

Figure 7.4 extends knowledge about planks by incorporating a list of conditions in description logic (shown in a white box). Conditions above the gray line are *sufficient* and *necessary*, while the ones below only *necessary*. These conditions not only augment the conceptualization of a plank, they also make explicit its properties and relationships. A couple of cases clarify this point. First, the condition *hasPlankToPlankJoinerySystem*, implies that planks can be joined together. This is a fact that can be grasped from the model (Figure 7.3) and the definition (Definition 7.1), but not so much from the photograph (Figure 7.1) and the illustration (Figure 7.2). This condition also makes explicit that there are different techniques for joining planks. Second, conditions *hasStrakes*, *hasWales*, and *hasGarboard*, indicate additional timbers that can be attached to planks. In addition, it can also be said that a **Plank** is a subclass of **HullComponent**.

The second example is **Frame**. Frame can be defined as: “A transverse timber, or line or assembly of timbers, that described the body shape of a vessel and to which the planking and ceiling were fastened.” This definition is richer than the one of plank because at least two concepts are made explicit: it is a transversal component, and names two other components to which it can be assembled: Ceiling and Planking. Figure 7.5 shows a photograph of a frame with the definition embedded. In this example textual and visual representations complement each other. The object expressed with words is enhanced with the graphic representation. Conversely, properties that are not explicit in the photograph are stated in the definition.

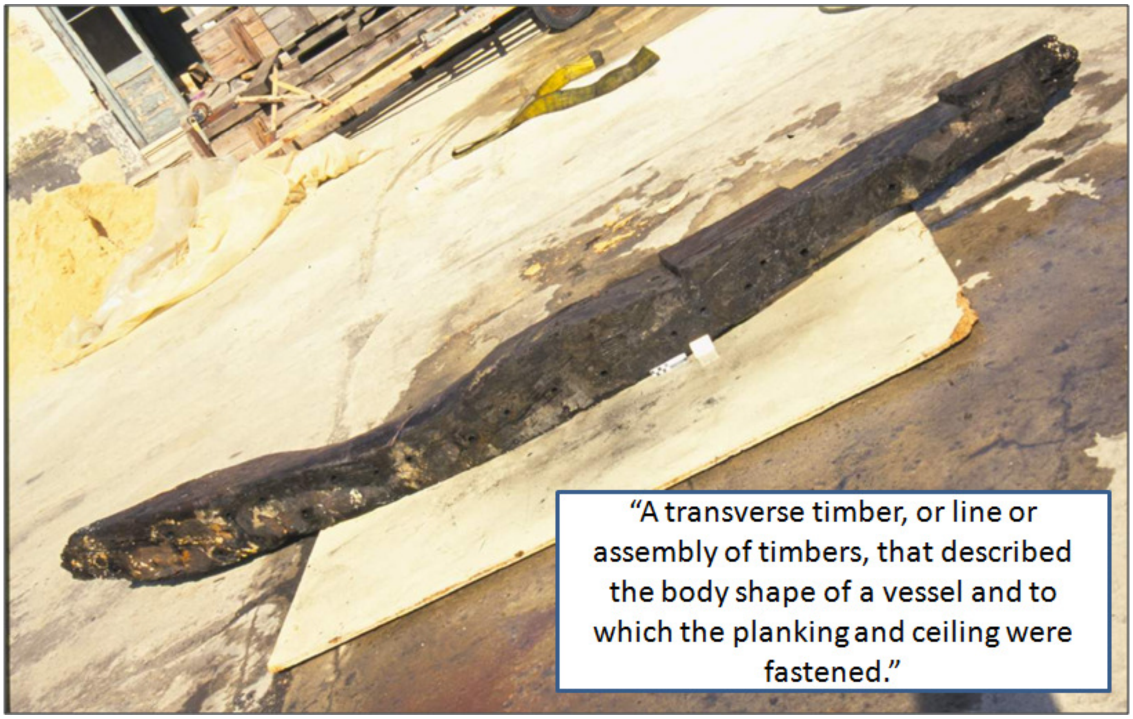


Figure 7.5. Photograph of a frame along with its definition. Photograph courtesy of CMAC.

An illustration of a master frame from Oliveira's *Livro da Fabrica das Naus* (c. 1580) is shown in Figure 7.6. The illustration offers a new perspective about frames. Original labels in the illustration correspond to proportions used for geometric descriptions in the text. The accompanying text provides detailed step by step assembly instructions.

A 3D model depicting frames in a wider context is shown in Figure 7.7. Comparing this model with the photograph in Figure 7.5 shows that presenting information in context improves the understanding about "things." In this case it is possible to visualize what is being expressed in the definition. Observing the model in Figure 7.7, it is possible to visualize **Frames** in context (as transversal components).

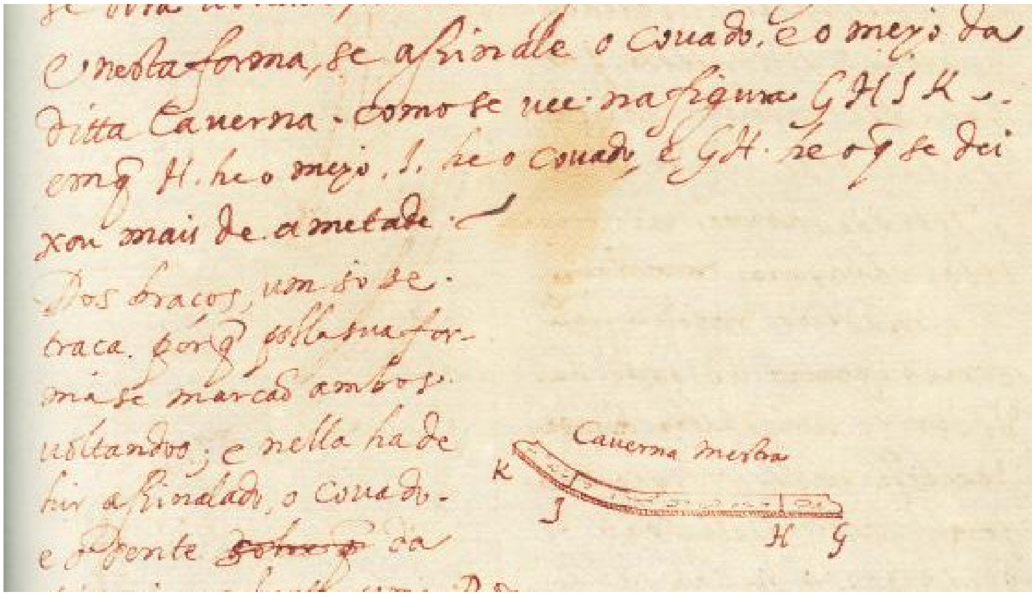


Figure 7.6. An illustration of a frame (caverna mestra in Portuguese) from Oliveira's *Livro da Fabrica das Naus* (c. 1580). Courtesy of NADL's on-line collection of shipbuilding treatises.

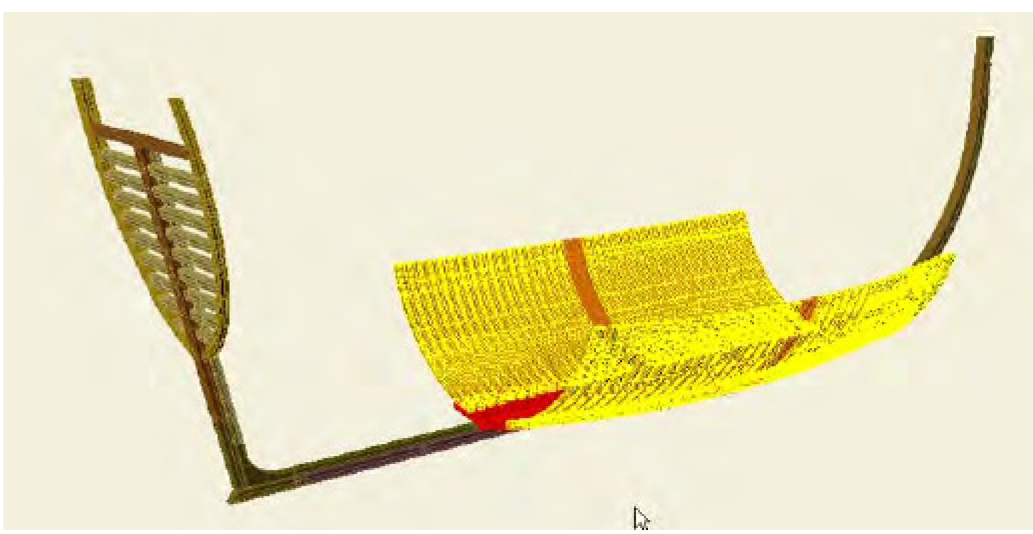


Figure 7.7. A 3D model showing frames in a ship. Model courtesy of Alex Hazzlet, CMAC.

Similar to Plank, there are still numerous conditions that are not known about Frames. Figure 7.8 shows the same photograph from Figure 7.5 along with the conditions from the ontology. *ontoShipDS's* conditions make explicit several additional facts about frames. For example that there are various framing techniques, that frames can be fastened, joined, or chocked; which in turn implies that there should be objects used in joining frames. In some cases **Rider Frames** (a Frame that goes on top of the ceiling) were also used along frames.

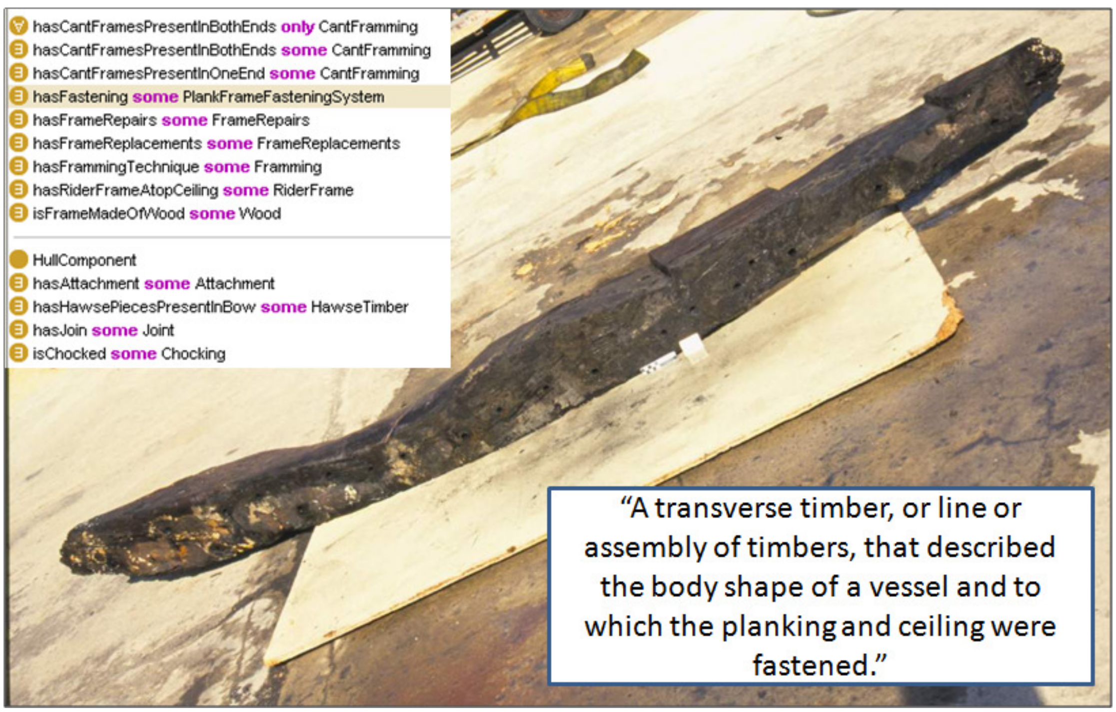


Figure 7.8. Photograph of a frame along with its definition and a list of rules describing its properties.

The preceding examples demonstrate how visual and textual abstractions complement each other. They also illustrate that conditions (from the ontology) not only enrich the understanding of objects, but also provide a formal definition of concepts.

7.2 Identifying Inconsistencies and Discovering Implied Relationships in the Original Model

Creating an ontology requires access to the information about the domain. Domain experts are the ones that analyze objects, their properties, relationships, and the classes they belong. The GeneOntology [Ashburner et al. 2000], for example, relies on a group of experts for maintaining an ontology about genes and their products. Similarly, *ontoShipDS* is based on the work of numerous nautical archaeologists. The findings of underwater excavations generate information that is catalogued and stored in paper-based archives and electronic files. Nonetheless, it was the work of Dick Steffy that for the first time attempted to use a relational database for enhancing both the storage of and access to the information (and preservation in the long term).

Although a relational database is not the best model for expressing the complexity of wooden ships, it offers a natural environment for representing objects and their relationships. It was indeed this database that served as a starting point for *ontoShipDS* (this was discussed in Section 6). Another outcome in the creation of the ontology from the existing information was to detect inconsistencies in the original representation. This can be attributed to the limitations of a database model in expressing richer relationships among objects.

This demonstrates how an ontology can be used in identifying inconsistencies in the formulation of a domain knowledge. In this section I discuss an example for detecting inconsistencies in the original representation. To accomplish this, screen shots of the original database tables are included along with my annotations, followed by various graphical representations in RDF-OWL.

V. **Hull Components** (a list of surviving hull timbers or timber fragments found on the wreck, indicated by a check mark if present)

1. ID (vessel ID)
2. K (keel or keel plank)
3. RK (rider keels)
4. SK (sister or bilge keels)
5. DW (deadwood)
6. SF (stem timbers, including aprons)
7. SA (sternpost timbers, including knees and heels)
8. SP (single planking)
9. DP (double planking)
10. BS (bilge strakes)
11. W (wales)

Figure 7.9. Partial list of the attributes in the original table hull components. From Dick Steffy's database.

Figure 7.9 depicts a partial list of the attributes in a table originally intended for storing hull components. Items 2 to 7, 10 and 11 correspond to components. On the other hand, items 8 and 9 indicate a technique (planking) rather than components. Further analysis suggested that single and double planking fit better under a different category (techniques). A **Plank** is a component; **Planking** on the other hand, is a technique with two instances: Single and Double.

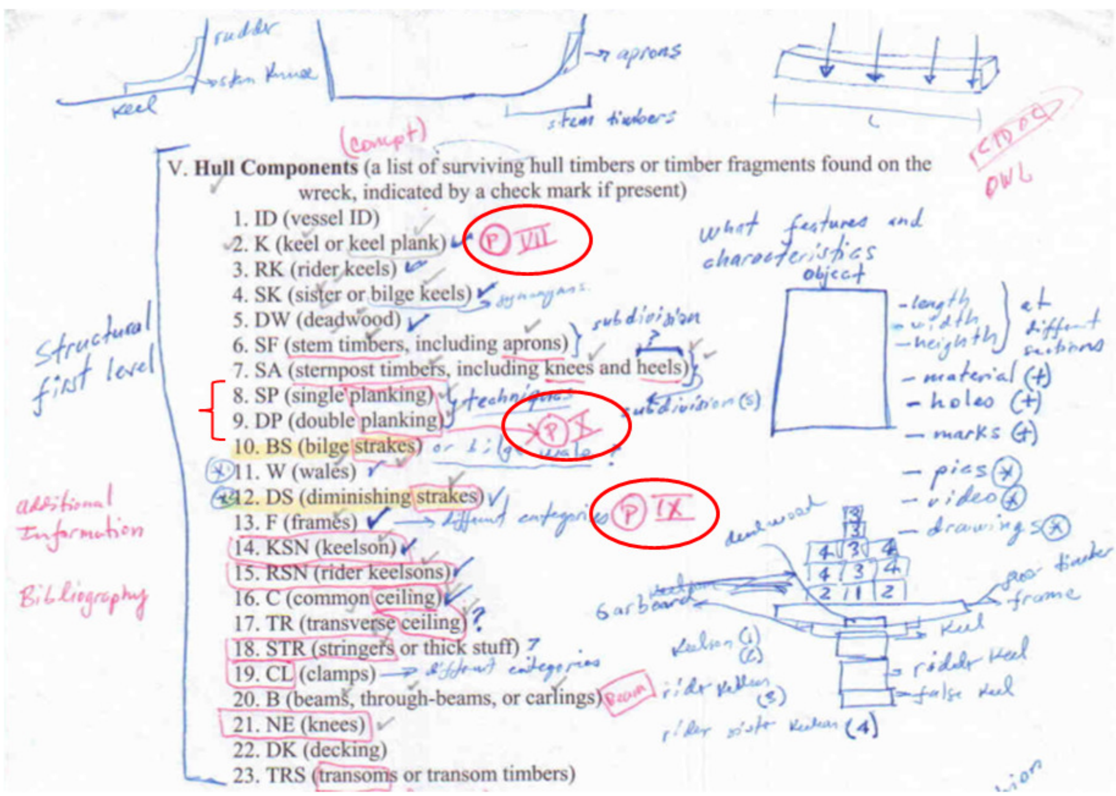
X. Planking

1. ID (vessel ID)
2. method (planking method; [PF] for planked on standing frames; [MT] for mortise-and-tenon joined seams; [LAP] for overlapping seams; [LIG] for ligatured seams; [COG] cogs and similar bottom-based forms; [OTH] for other; [UNK] for unknown. NOTE: precede with D if double-planked; T if triple-planked}
3. strakes (number of strakes or planks in evidence; follow number with P if individual planking pieces only)
4. scarfs (types of scarfs or butts used – [D] for diagonal; [S] for S scarf; [Z] for three-planed; [B] for butt join; [V] for vertical scarf; [O] for other)
5. garboard (garboard thickness at keel, in cm)
6. plank (common planking thickness, in cm)
7. wales (number of wales in evidence)
8. plan (are wales belted [B] or separated by spacing planks [S}?)
9. main (main wale thickness, in cm)
10. dimin (are there diminishing strakes? check if yes)
11. pl rep (is there evidence for planking repairs or replacements? check if yes)

Figure 7.10. Original structure of table planking. From Dick Steffy's database.

A separate table named Planking (Figure 7.10) was used to describe **Planks**. The table's name in itself is ambiguous because **Planking** can be a verb. Item 2 indicates various planking methods (or techniques). Each one of them can be applied to double or single-planking. Items 3, 4, 5, and 7 suggest that a **Plank** is related to **Strakes**, **Scarfs**, **Garboards**, and **Wales**. Item 4 in turn indicates various types of **Scarfs**. Items 3 and 10 belong to the same concept: **Strakes.** Items 7 and 9 belong to **Wales**. Items 6, 8, and 11 represent properties of **Planks**.

Figure 7.11 shows my own annotations during the transformation and disambiguation of the original database design. Items 8 and 9 are marked as techniques. They were merged under the class **Plank**. In addition, the class **Planking Pattern** was created and placed under **Technique** along with three subclasses: **Single**, **Double,** and **Triple Planking**. Items 2 and 13 had originally their own tables Keel/Keel Plank and Planking respectively. They were renamed as **Keel** and **Plank**, and classified under **Hull Component**. The goal in this process was to separate objects, characteristics, and techniques. It also allowed the elimination of redundancies, and classification of items according to their properties.



V. **Hull Components** (a list of surviving hull timbers or timber fragments found on the wreck, indicated by a check mark if present)

1. ID (vessel ID)
2. K (keel or keel plank)
3. RK (rider keels)
4. SK (sister or bilge keels)
5. DW (deadwood)
6. SF (stem timbers, including aprons)
7. SA (sternpost timbers, including knees and heels)
8. SP (single planking)
9. DP (double planking)
10. BS (bilge strakes)
11. W (wales)
12. DS (diminishing strakes)
13. F (frames)
14. KSN (keelson)
15. RSN (rider keelsons)
16. C (common ceiling)
17. TR (transverse ceiling)
18. STR (stringers or thick stuff)
19. CL (clamps)
20. B (beams, through-beams, or carlings)
21. NE (knees)
22. DK (decking)
23. TRS (transoms or transom timbers)

Figure 7.11. My own annotations in the transformation and disambiguation of the table hull component.

The next step was to create properties for expressing how these components were related. These properties express relationships among Plank and other classes (see Figure 7.12). For example: *hasGarboard* states that a **Plank** is related to **Garboard**, *hasPlankingPattern*, indicates that a **Plank** can follow diverse **PlankingPatterns**. **Planks** can be placed following various techniques, this is indicated by **Planking**. We can also infer that **Planking** is a technique because it is a subclass of **Technique**. Similarly, other properties can be discovered. For instance the last property in the list *playsAStructuralRole* can be used to find all components that have a structural role in the ship. Although this is a simple example, it illustrates how the ontology offers a richer description of concepts, how it can be used in the disambiguation of terms and concepts, and for discovering implied relationships.

Figure 7.12. A screen shot generated in Protégé-OWL’s user interface
depicting the properties of plank.

Although not exempt from ambiguity, tables depicted in figures 7.9 and 7.10 express (at a higher level) knowledge about certain concepts. They capture an initial idea about planks. So how were inconsistencies and ambiguity in these concepts removed during the creation of *ontoShipDS*? The top level corresponds to the most general concepts: **Component**, **Technique, Ship, Property, Concept,** and **Object** (see Figure 7.13). **Component** was broken down into **Hull Component**, **Auxiliary Component**, and **Internal Structure** (Figure 7.14).

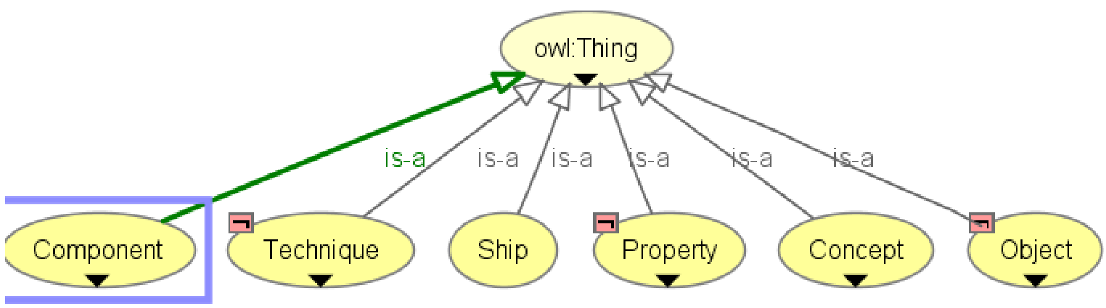


Figure 7.13. An OWL Viz-generated graph depicting the high level classes in *ontoShipDS*.

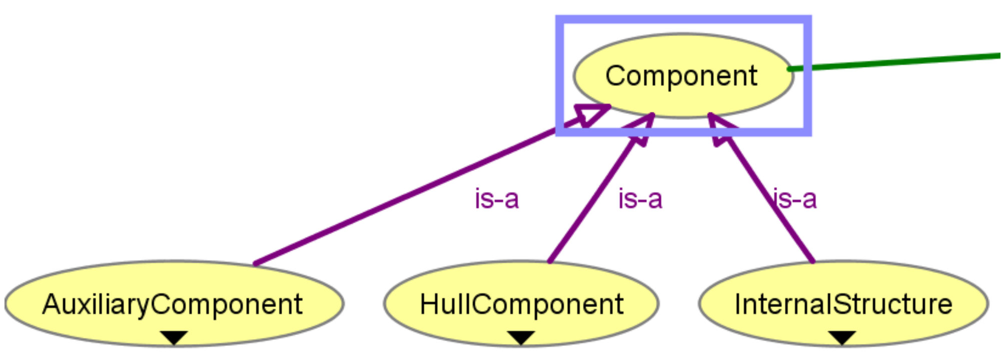


Figure 7.14. An OWL Viz-generated graph depicting classes that belong to the super class component.

Another consequence in the creation of *ontoShipDS* was the possibility of adding structure to the implied relationships. As shown in this example, the relationship hasPlanking(Plank, Planking) enabled restructuring the original representation in two ways, first by merging some terms, and second by expanding and re-classifying others.

7.3 Capturing the Knowledge from Archaeologists

A major advantage in the creation of *ontoShipDS* is capturing the knowledge from archaeologists. Before its creation there was no formalization about wooden ships. This knowledge was scattered in publications, field notes, electronic files, and personal communications. Capturing this knowledge helps to preserve archaeological information. An archaeologist formalizing a given discovery in an ontology facilitates its accessibility by members of the community. It also makes them aware not only of new discoveries, but also on how they relate to the entire corpus, what existing understanding is challenged, and the new knowledge created.

An ontology is also a tool for disseminating the knowledge in nautical archaeology. A formalization of wooden ships facilitates the communication not only among scholars and researchers, but also to the general public. Given the degree of specialization of this discipline, the use of unknown terms in various languages makes the dissemination of information among non experts more difficult (even experts are not exempt from this problem). The ontology helps by providing a mechanism for expressing the knowledge in the domain in multiple languages.

This improvement was evident when transforming Steffy's database. His relational model was already an attempt in formalizing the conceptualization of wooden ships. Analyzing his design and reclassifying concepts shed light on how wooden vessels were perceived, what concepts and properties were important, and what real examples documented that knowledge. Because *ontoShipDS* is expressed in RDF-OWL, it enables portability, making it possible to be used in different environments.

Although the mere capture of knowledge from nautical archaeologists is in itself a great advancement, it is the dissemination and preservation that makes it even more valuable. A group that will benefit from this ontological representation is nautical archaeology students. These students are less specialized than archaeologists, but much more than the general public. What they are trying to accomplish is to understand ship construction. A personal experience serves to illustrate this. A few years ago when I enrolled in the class Treatises and Books on Shipbuilding, offered by the Nautical Archaeology program at Texas A&M University, we had to access various shipbuilding treatises and discuss their contents. The goal was to understand diverse shipbuilding conceptualizations and techniques. The lack of a formal representation made it more complex and difficult, although in my case, not having a background in nautical archaeology could accentuate this difficulty. In retrospect, I believe that an ontology such as *ontoShipDS* would have been extremely helpful.

7.4 Reasoning About Definitions

Another application of *ontoShipDS* is to be used for reasoning about definitions. For example, let us define the properties *usesJoiningObject* and *isAJoiningTechnique* to indicate that a joining technique requires certain objects, and that a given technique can only be applied to a certain ship components.

Let us consider a hypothetical scenario, if we state that a **Tenon** can only be used for joining **Planks** to **Planks**, and that a technique that uses **Tenons** can only be employed on single-layered hulls. A reasoner using this ontology can tell us that if **Tenons** are found in an excavation, the construction technique used on the hull has to be single-layered. Conversely, the reasoner can suggest that a hypothesis is false, or at least has to be revised. For example, an archaeologist who has previously hypothesized that recovered planks from a shipwreck under study indicate that the ship was a double-layered hull, may need to revise the hypothesis based on new evidence recovered (**Tenons** for example). In a general sense the ontology can be used to prove or reject hypotheses.

Note that it is possible that this new discovery indicates that Tenons can also be employed in double-planked hulls. In this case, it is not that the reasoner failed, rather that with the knowledge available at that time, the inference indicated that such fact was not valid. However, if the new discovery holds true, the ontology has to be adjusted to reflect the new discovery.

7.5 Suggesting Facts and Properties

From the digital libraries standpoint, the ontology can be used as bridge for connecting resources and for expanding their definitions. The new type of queries that can be used are not limited to keywords only. They can be expanded to include relationships and properties. Similarly, an ontology offers various ways for browsing a collection. Exploration and navigation can start from a particular object. Following that object's relationships or properties can take users to other objects that were probably not intended from the beginning, but enrich the experience. This scenario has some similarities with a suggestion system in on-line shopping sites. For example looking at an item X, users are presented with suggestions about similar or related objects. These objects are presented based on categories or relationships that can be expressed as an ontology.

7.6 Expanding Knowledge from Fractional Information

Archaeology is a field where evidence is often damaged or incomplete, or for which no related artifacts are found. *ontoShipDS* can be used to expand the knowledge about an object under study. For instance, the diameter of a hole in a timber could indicate the dimensions of the joining objects. Yet, the type of joining objects used might not be known. The ontology can suggest whether nails or spikes were used for that particular component, and how they were used.

Another case would be one where patterns of holes and cuts in a timber might hint at the dimensions of the other assembled components, even though none were found. In a broader sense, the ontology helps to expand facts and properties of objects. This can help to reduce the probability of leaving out information, and to make explicit what facts are important to inspect and analyze. In a post excavation scenario, the ontology can be coupled with photographs or drawings documenting archaeological evidence, enhancing and facilitating their understanding.

Finally, an ontology can help to compare an implied relationship in a particular model against explicit data found in a real excavation. For example, common knowledge expressed in the ontology states that ships from a particular geographical region or in a given period of time had frames attached to wales. However, evidence from the excavation suggests that wales could not be attached to frames.

8. EVALUATION AND DISCUSSION

My goal in this dissertation is to improve the contextualization of terms in nautical technical texts by means of linking information in a more meaningful way, and ultimately assisting archaeologists in the reconstruction of sunken ships. Better contextualized terms allow the presentation of information in a more relevant fashion. The proposed approach is based on exploiting information from two external sources: *glosShip*, a multilingual glossary of nautical terms, and *ontoShipDS*, an RDF-OWL ontology of wooden ships.

In order to determine the effect of these external sources for improving contextualization, an evaluation was conducted. Three methods for improving the ranking of query results were compared. The method "text co-occurrence" uses co-occurrences of glossary terms in the texts; results are ranked based on term frequency. Terms that co-occur more frequently are assigned a higher weight; hence paragraphs that include those terms receive higher scores and will be ranked at the top.

Method "definition co-occurrence" uses co-occurrences of glossary terms within their definitions. Because it is not possible to automatically determine their importance, all co-occurrences receive the same weight. This weight is used for ranking paragraphs that include a given search term.

The third method, "ontology-ranked" exploits information (such as techniques, properties, and relationships) defined in the ontology for determining term weights in the two methods aforementioned. Query results are then ranked based on different perspectives in which ship components can be explored. If the goal is to see how a component is assembled, information about techniques from the glossary is employed. If measurements and materials of a given component are the focus of analysis, information about properties from the glossary is used. Similarly, determining what components are attached to a given ship part requires the use of information about their relationships.

8.1 Evaluation Design

The materials used in the evaluation included texts pertaining to shipbuilding treatises in various languages. The Spanish collection included three *Ordenanzas de fábricas de navios* published in 1607, 1613, and 1618. The Portuguese collection comprised the following treatises: *Coriosidades* de Gonçallo de Sousa (c. 1630), *Livro Nautico* (1575-1625), *Liuro da Fabrica das Naus* (c. 1580), and *Relacao*. Materials in English include descriptions of about a hundred underwater excavations compiled by Mr. Dick Steffy. Two Venetian sources, the *Trombetta Manuscript* (c. 1444), and Nicolo e Hieronimo Secula's *Modo di far galee grossi e sottili* (c. 1556), completed the written sources. Due to copyright restrictions on some of the transcriptions, they were used internally for evaluation purposes only.

The multilingual glossary contains the following number of terms: English 3,210; Portuguese 481; Spanish 419; Dutch 2,475; Italian 32; Latin 74; French 873; Catalan 131; German 87; Danish 80; Greek 208; and Venetian 6. The numbers vary depending on accessibility to sources (dictionaries, encyclopedias, or glossaries in other languages), and availability of experts for editing the translations.

For indexing purposes, each line in a text was considered a document. This level of granularity was determined, based on the contextual significance of each individual term. Hence, using a treatise or even a page as document was not meaningful.

A collection of 1530 tagged images in various languages was used for including visual representations of ships in the evaluation. These images contain illustrations and texts from the treatises in the NADL collection, along with approximately 600 photographs and drawings pertaining to the *Pepper Wreck* excavation. In addition, 65 images of a 3D model of a Portuguese vessel created in Maya (courtesy of Dr. Alex Hazlett) were used.

The expansion algorithm was based on exploiting information from the glossary in various languages. A full-text index was created using Lucene. Due to meaningful terms in nautical archaeology included in Lucene's built-in stop word lists (see appendix C), customized lists were used. In order to have a diverse set of terms for the evaluation, they were selected from various categories in the ontology.

A preliminary step was to determine relevant documents. Relevance is a key component in measuring performance in information retrieval. In TREC, a document is considered to be relevant based on relevance judgments given by human experts, using a pooling technique. In this dissertation's evaluation, relevance of text snippets was determined according to various contexts: a) spatial location, b) relationship to other components, c) role of the component, and d) assembling techniques.

8.2 Weighting Functions

Ranking documents for improving contextualization requires a function (or functions) that describe how weights are calculated. In this dissertation, three weighting functions were defined, one for each method.

The weighting function for the method "text co-occurrence" is defined as follows: the weight of term T_i is equal to the number (frequency) of co-occurrences of that term (at the cut-off point n) divided by the total number of co-occurrences of all the terms at the cut-off point n. Therefore, high frequency terms will receive higher weights.

$$w(T_i) = \frac{f(T_i)}{\sum_n^{j=1} f(T_j)} \qquad \text{Equation (8.1)}$$

The weighting function for the method "definition co-occurrence" is defined as follows: any term whose definition includes term T_i receives a weight of 0.75. A term in the definition of term T_i that has entry in the glossary receives a weight of 1. The idea of this calculation is that terms included in the definition of another term T_i are more likely to have a strong relationship. Conversely, terms whose definition includes term T_i are also related, but probably to a lesser degree, hence a lower weight.

$$\mathrm{w(T_i)} = .\left\{ \begin{array}{lllll} if & T_i & in & Def(T_{i \neq j}) & = 0.75 \\ if & T_i & in & Def(T_i) & = 1.0 \\ \text{else} & & & & = 0 \end{array} \right\} \qquad \text{Equation (8.2)}$$

The method "ontology-ranked" assigns a weight of 1 to terms that are related to a given term T_i based on some conditions (functional, spatial, or properties) according to the ontology Θ. Although this is a simplified version, related terms receive higher weights, ranking documents where they appear, at the top of the results.

$$w(T_i) = \begin{Bmatrix} if \quad T_i \quad in \quad \Theta & = & 1.0 \\ \\ else & = & 0 \end{Bmatrix} \qquad \text{Equation (8.3)}$$

8.3 Characteristics of the Glossary

Since information from the glossary was used in the expansion algorithm, this section includes some quantitative information about its contents. English was chosen as reference language because it has the largest number of terms. Lexical expansion is the number of composite terms in the glossary that contain another term. Term **false keel** contains the word **keel**; therefore, it is used in the lexical expansion of keel. Figure 8.1 shows a partial list of the lexical expansion for the word **keel**.

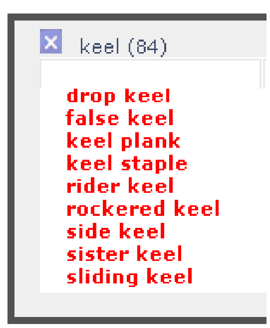


Figure 8.1. Partial lexical expansion of the term Keel.

Figure 8.2 shows the distribution of lexical expansion in the glossary. Out of 2,700 terms, the top three categories are: 492 have one expansion term; 251, 2 terms; and 259, 3 and 4 terms. Although not used in this evaluation, lexical expansion can be used in ranking search results. Thus, entries with more lexical expansion terms will increase term weight during retrieval.

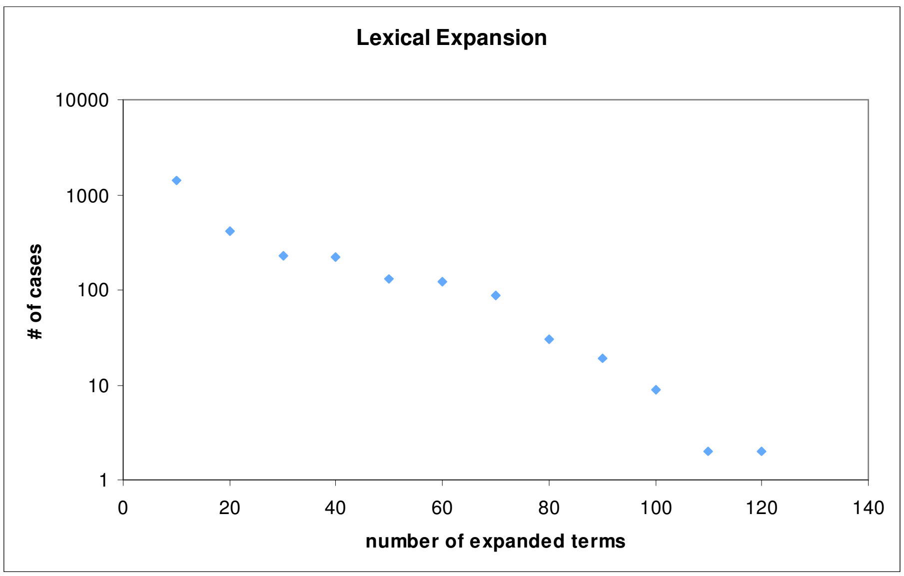


Figure 8.2. Distribution of lexical expansion per terms in the glossary.

Two other properties from the glossary that are used in this evaluation are synonyms and spellings. Query terms are first expanded into their corresponding synonyms and spellings, which enables the retrieval of additional information. Figure 8.3 depicts the distribution of synonyms and spellings in the glossary. The scale in the graph is logarithmic for presentation purposes. Despite the relatively low number of terms with synonyms or spellings (14% and 6%), they increased the retrieval of textual snippets.

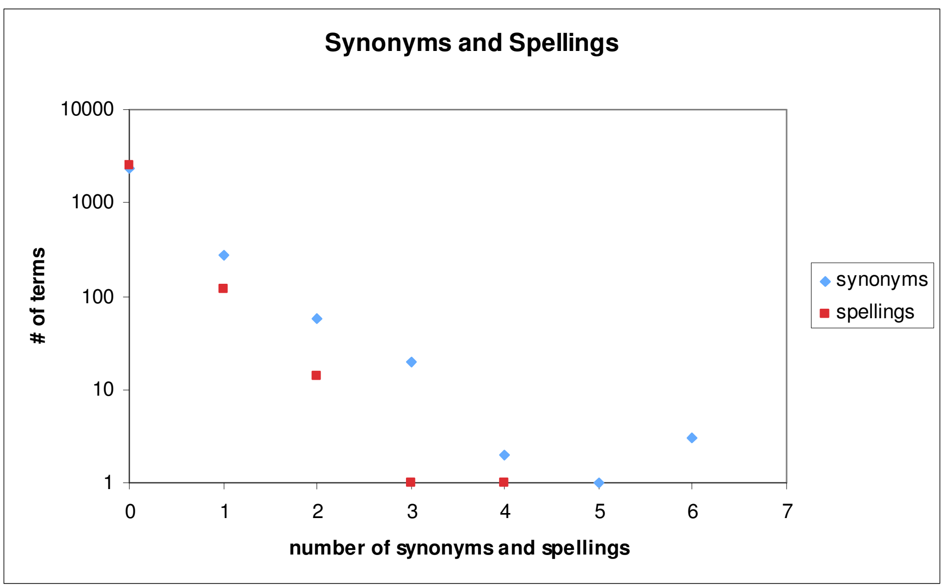


Figure 8.3. Distribution of synonyms and spellings in the glossary.

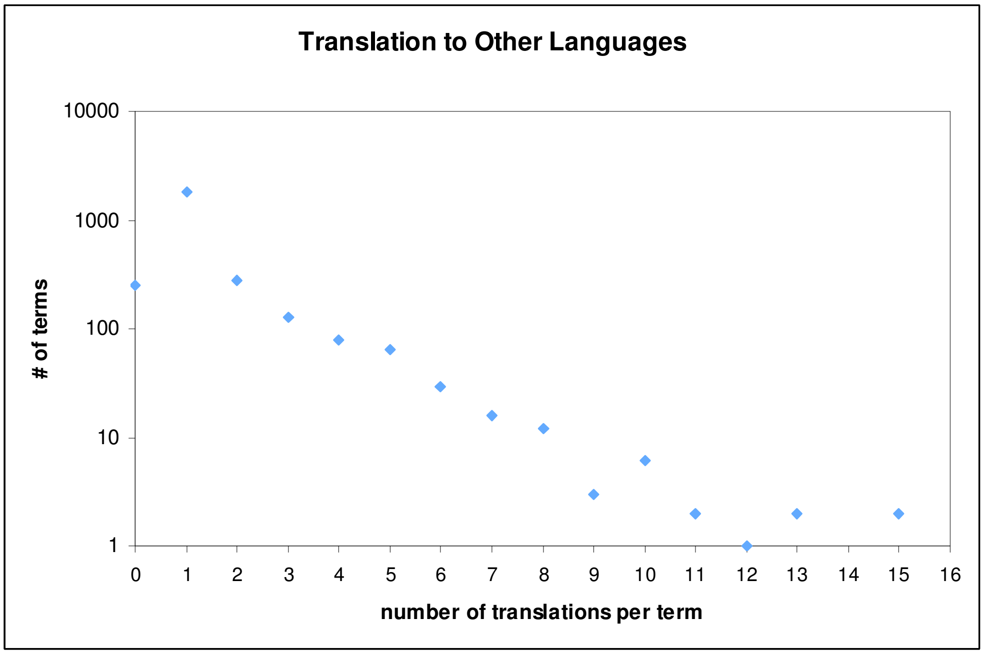


Figure 8.4. Distribution of translations to other languages.

Given the multilingual nature of the glossary, terms translated into other languages increase information retrieval. Figure 8.4 shows the distribution of translations. Terms with 2 and 4 translations accounted for 78% and 7.6% respectively. A total of 252 terms (9.3%) did not have translations at all. In this context, number of translations is, for a given term, the number of translated terms to other languages; and not the number of languages a term was translated into. This explains why even though the glossary has twelve languages, there were a few entries with fourteen or more translations.

Figure 8.5 shows a partial depiction of the graphic interface of the prototype listing snippets in Portuguese (escarva) and Spanish (junta) expanded from the original

English query (scarf). Due to stemming, some false positives (words enclosed in red rectangles) are also retrieved. In Spanish, the translation of scarf (*junta*) is stemmed to "junt," while the Portuguese translation (*escarva*) to "escarv."

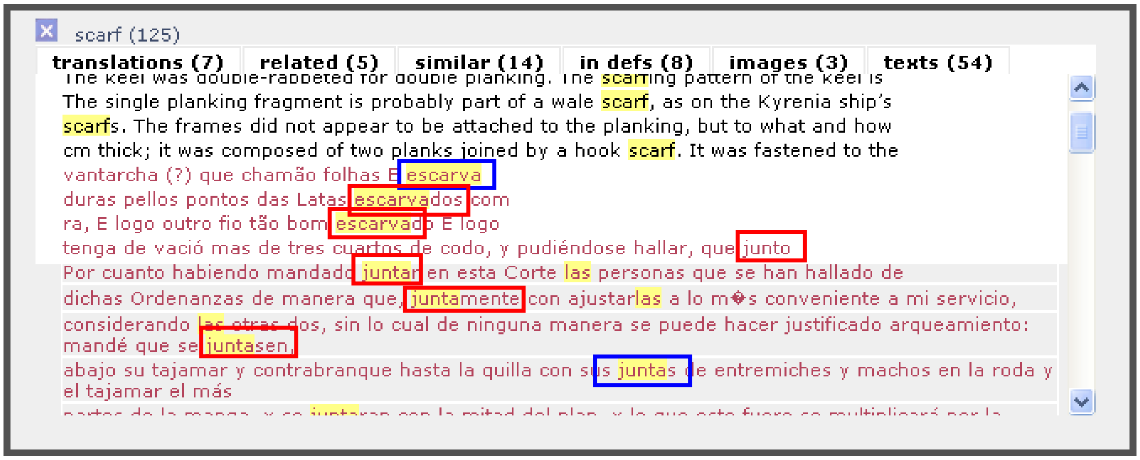


Figure 8.5. Partial list of results in various languages including false positives.

The expansion algorithm also makes use of definitions for improving information retrieval. The goal is to discover conceptual relationships, using terms within the definition that have an entry in the glossary. This information can be used for weighting retrieved results, and ranking them according to a topic in particular. Out of the terms that have a definition, the most significant are: with 3 terms 50%; with 6 terms, 21%; with 9 terms, 13%, and with 12 terms, 7% (see Figure 8.6).

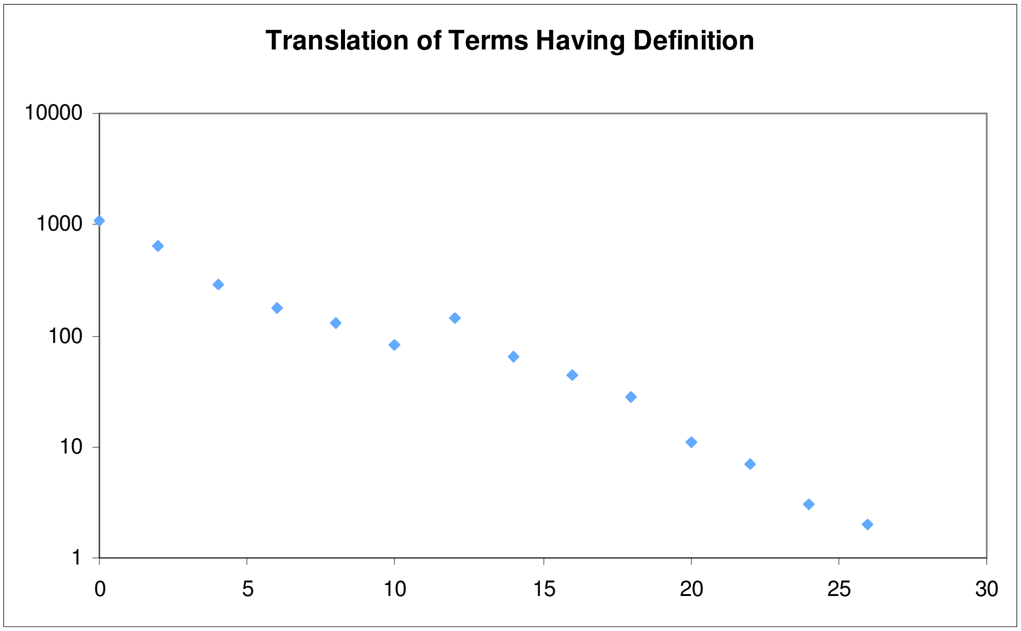


Figure 8.6. Distribution of terms according to the number of references in their definition.

8.4 Experiment 1: Using Corpus Co-occurrence for Contextualization

In the evaluation of the "text co-occurrence" method, term co-occurrence in the texts was calculated. Co-occurrence was limited to words with an entry in the glossary, since they are meaningful in nautical archaeology. In information retrieval, different window sizes have been proposed for determining co-occurrence [Vechtomova and Wang 2006]. In this dissertation paragraphs were used as context window. For each term, frequency and average distance were calculated. Frequency is the number of occurrences; distance on the other hand, is the number of words separating one term from the other within the established window.

A sample of terms was used to analyze the behavior of term co-occurrence in various languages. The following tables contain results of co-occurrences ranked by frequency derived from texts in the available languages. Frequency was chosen for ranking purposes. Average distance is rounded to its upper value. Term identifiers are included (first column in each language labeled id) to facilitate the comparison among languages.

The expansion algorithm employed in this evaluation was applied to the entire glossary in all languages and the whole collection of texts. With thousands of terms in the glossary, the number of languages of the texts, and the possible combinations, it is not surprising that the file containing the results had almost half a million lines. With this scenario, a comprehensive analysis in the scope of this dissertation would be impractical. Therefore, a sample of terms was selected to evaluate the proposed approach. These terms were chosen because they are very important structural elements of most ships (F. Castro, personal communication, March 29, 2010). The results of the evaluation are discussed in the following paragraphs. Grayed cells in the tables correspond to the id of terms that rank in the top-n across languages. The cut-off *n* was set at 15. This value provides a good window of term co-occurrence. Increasing the cut-off value allows the inclusion of more terms, but decreases frequency; hence, reducing their importance. In some instances, such as Table 8.3, the cut-off value differs for the same term in different languages. This means that there are fewer or more co-occurrences in a particular language than in others at a given cut-off value. This does not affect the results of the experiment because co-occurrences of the same term in different languages are not compared against each other.

Results of *keel* are shown in Table 8.1. Although with different rankings, *keel* in English, Portuguese, and Spanish share 2 terms in the top 15 co-occurrences: *stem post* (138) and *sternpost* (143). *Floor timber* (56) and *bottom* (17) are top ranked in English and Portuguese, while *proa* (19), *popa* (139), *coberta* (46), and *gio* (157)—bow, stern, deck, and tuck—are top-ranked in both Portuguese and Spanish. These results suggest semantic relationships among the terms. The word *bottom* (17) indicates a spatial characteristic (its location) of the keel (at the bottom of a ship), which can be validated with its definition from the glossary:

> **bottom**: The part of a hull between the bilges, including the *keel*.

Similarly, *sternpost* (143)*, stem post* (138)*,* and *floor timber* (56) are related to *keel*, this is confirmed by their definitions (from the glossary), as follows:

> **sternpost**: A heavy timber on the *keel* end usually inclined slightly aft on which planking terminates and any stern frame is constructed, and on which the rudder is hung.
>
> **stem post**: A vertical or upward curving timber or assembly of timbers scarfed to the *keel* or central plank at its lower end, into which the two sides of the bow were joined.
>
> **floor timber**: The central element of a frame. The timber that straddles the *keel* or keelson(s) upon which the remainder of the frame is build on.

Analyzing Table 8.1 shows some terms top ranked in more than one language (gray cells in the id column). In contrast, other terms are top ranked in only one or two languages (white cells in the id column). Whether terms are top ranked in one or more languages does not reduce their usefulness for relevance feedback. In fact, they still provide additional contextual information for query ranking.

Table 8.1. Results of the top 15 co-occurrence with keel in English, Portuguese, and Spanish texts.

English				Portuguese				Spanish			
id	Term	avg. dist.	freq	id	term	avg. dist.	freq	id	term	avg. dist.	freq
56	floor timber	8	27	138	roda	33	116	19	proa	96	232
138	stem post	11	12	19	proa	48	108	139	rasel de popa	86	194
66	garboard	5	8	143	cadaste	28	94	46	cubierta	45	182
144	strake	6	7	139	popa	53	70	49	puntal	68	180
79	hull	7	6	124	ragel	52	69	32	esloria	32	136
457	molded	13	6	98	mastro	61	54	157	yugo	61	125
483	planking	6	6	46	coberta	51	52	8	manga	65	110
2503	top	3	6	86	sobrequilha	8	40	173	horcate	121	108
546	side	3	5	250	largura	15	38	55	plan	49	80
460	mortise	16	4	1997	lancamento	51	26	47	bao	98	72
1873	pine	5	4	157	gio	9	25	138	branque	29	52
114	rabbet	3	4	56	caverna	51	22	143	codaste	56	40
17	bottom	10	3	41	dormente	89	22	584	segunda	102	28
387	heel	3	3	8	Boca	25	20	119	astilla	69	24
143	sternpost	2	3	17	fundo	13	15	130	buque	43	24

Analysis of *stem post* shows that *stern* (139), *bow* (19), *keel* (84), and *sternpost* (143) co-occur in the top 12 in three languages, see Table 8.2. In contrast, *deck* (46) co-occurs only in Portuguese and Spanish. From the semantic standpoint, the relationship with the term *bow* suggests a reference to the front of the ship; hence, it can be used for narrowing down query results based on spatial properties of the ship components. Similarly, in the case of an image depicting a section of a ship, it can be used to better indicate the location of the *stem post*.

Table 8.2. Results of the top 12 co-occurrences with stem post in English, Spanish, and Portuguese texts.

English				Portuguese					Spanish			
id	term	avg. dist.	freq	id	term	avg. dist.	freq		id	term	avg. dist.	freq
2503	top	4	12	19	proa	36	256		46	cubierta	61	112
276	chock	7	6	84	quilha	31	104		19	proa	25	98
139	stern	4	6	143	cadaste	29	98		138	branque	33	70
1180	elm	8	3	139	popa	31	87		139	rasel de popa	33	58
2486	timber	7	3	46	coberta	73	74		32	esloria	23	56
19	bow	7	2	250	largura	67	50		7	espolon	53	38
79	hull	9	2	1997	lancamento	37	49		157	yugo	39	35
1267	fore	13	1	8	Boca	26	47		135	lanzamiento	14	32
84	keel	11	1	98	mastro	81	46		84	quilla	30	27
460	mortise	5	1	124	ragel	79	36		173	horcate	83	24
143	sternpost	2	1	121	leme	14	32		48	tablón de cubierta	42	21
									143	codaste	41	20

Results for *transom* are depicted in Table 8.3. Both Portuguese and Spanish share various terms in the top 15. *Popa* and *rasel de popa* (139, stern in English). C*adaste* and *codaste* (143, sternpost in English), strongly suggest that transom is a ship component located at the rear (stern) of the ship. C*oberta* and *cubierta* (46, deck in English) indicate that the transom is placed at a high distance, close to the deck. This is confirmed by the definition of **transom** from the glossary: "A flat termination to a stern. Above the water line." Co-occurrences in English are not included because none were found in the English texts.

Table 8.3. Results of co-occurrences with transom in Spanish and Portuguese texts.

Portuguese				Spanish			
id	Term	avg. dist.	freq	id	Term	avg. dist.	freq
143	Cadaste	30	26	46	cubierta	46	170
139	Popa	39	24	139	rasel de popa	14	108
46	Coberta	32	20	19	Proa	20	88
45	Abóbada	25	18	32	Esloria	48	64
19	Proa	64	16	47	Bao	59	42
29	barras do cabrestante	7	12	8	Manga	33	33
98	Mastro	54	12	49	Puntal	40	28
8	Boca	20	11	85	Quilla	48	25
250	Largura	12	10	98	Mástil	8	24
121	Leme	15	10	138	branque	35	16
138	Roda	33	10	157	Yugo	11	10
164	Cinta	44	9	143	codaste	31	7
84	Quilha	18	8	48	Tablón de cubierta	50	6
455	Forma	39	7	125	Junta	3	5
204	Grade	24	7	124	Rasel	10	4

8.5 Experiment 2: Using Glossary Derived Co-occurrences for Contextualization

In the previous section, semantic relations were derived from term co-occurrence in the texts. The second experiment is aimed at identifying term co-occurrence from the definitions in the glossary. Similar to co-occurrences from the corpus, I hypothesize that definitions can also indicate relationships among ship components.

Co-occurrence from the glossary was calculated as follows: first, for a given term, identify all words in its definition that have entry in the glossary. Second, find all terms in the glossary whose definitions include the given term. This calculation is slightly different from the one used in the "text co-occurrence" method. Neither distance nor frequency was estimated, because the length of the definitions is not significant.

In order to compare results, definitions of the terms used in Section 8.2 were selected. For *keel*, Table 8.4 lists terms in its definition that have entries in the glossary. Table 8.5 shows terms in any language whose definitions include *keel*. In both tables, the column labeled (r, ad, f), indicates the ranking, average distance, and frequency values calculated in experiment 1. The first value is the rank number, the second is the average distance, and the third the frequency. Those terms not included in the ranking are indicated as N/A. Similarly, Tables 8.6 and 8.7 list the terms in English and Portuguese whose definitions include *keel* and *quilla* respectively.

From these results, *floor timber*, *stem post*, and *garboard* obtained rankings 1, 2, and 3 from co-occurrence in the English texts. In contrast, *cadaste*, was ranked 3 from co-occurrence in Portuguese texts.

Table 8.4. Terms in the definition of keel with entry in the glossary.

English	**(r, ad, f)**
timber	(16, 12, 3)
hull	(5, 7, 6)
keel	N/A
longitudinal	(19, 10, 2)
molded	(6, 13, 6)
ship	N/A
sided	(9, 3, 5)

Table 8.5. Terms in the glossary whose definitions include the term keel.

English	**(r, ad, f)**	**Portuguese**	**(r, ad, f)**
Bottom	(13,10,3)	alefriz	(54,15,1)
floor timber	(1,8,27)	cadaste	(3,28,94)
Futtock	(27,10,1)	caverna	(12,50,22)
garboard	(3,5,8)	largura	(9,16,38)
Stern	N/A		
Rabbet	(12,3,4)		
Heel	(14,3,3)		
Keelson	(28,2,1)		
Skeg	(33,13,1)		
stem post	(2,11,12)		
Sternpost	(15,2,3)		

Terms in the definition of *stem post* with entry in the glossary are listed in Table 8.6. *Timber*, *bow*, and *keel* are terms that co-occur with *stem post* in English with rankings 5, 6, and 9. *Plank*, however, does not co-occur in the texts. Conversely, from the Portuguese definition, the terms *proa* and *quilha*, co-occur in the Portuguese texts with rankings 1 and 2, while *forma* and *navio* with rankings 10 and 18.

Table 8.7 shows terms in any language whose definitions include *transom*. They correspond to *bow* in English; *arco* and *beque* in Portuguese, with rankings 7, 51, and 31. Unlike the case of *keel*, results for *transom* do not suggest a strong correlation between corpus co-occurrence and definitions in the glossary. However, the glossary has a definition for *transom* in Catalan, which can be used for query expansion with materials in Catalan.

Table 8.6. Terms in the definition of stem post with entry in the glossary in English and Portuguese.

English	**(r, ad, f)**
Bow	(6,7,2)
Keel	(9,11,1)
Plank	N/A
Timber	(5,7,3)

Portuguese	**(r, ad, f)**
Forma	(18,38,13)
Navio	(10,30,23)
Proa	(1,31,256)
Quilha	(2,31,104)

Table 8.7. Terms in the definition of transom with entry in the glossary in English, Portuguese, and Catalan.

English	**(r, ad, f)**	**Portuguese**	**(r, ad, f)**	**Catalan**	**(r, ad, f)**
bow	(7, 2, 250)	Arco	(51, 4, 2)	Bombes	N/A
		Beque	N/A	Cavilles	N/A
		Buçarda	(31, 93, 5)	corbatons de la paramola	N/A
		Capelo	N/A		
		coral da popa	N/A		
		coral da proa	N/A		
		delgado da proa	N/A		
		lançamento da roda	N/A		

The term *transom* showed a smaller number of derived terms from the definitions. In Portuguese there was one occurrence, the term *popa*, with values (2, 39, 24) from the "text co-occurrence" method (discussed in Section 8.2). Spanish had also one occurrence, the term *codaste,* with values (12, 31, 7). Of these values, *popa* seems to indicate a strong correlation, since it is ranked 2. Validating this result with the definition of transom, it can be established that *transom* is in fact located at the stern of a ship.

8.6 Experiment 3: Using *ontoShipDS* for Term Contextualization

Since the ontology describes objects, properties, and their relationships, it can be used to enhance the context of terms. Results of the experiments in Sections 8.2 and 8.3 serve to illustrate this point. As stated earlier, co-occurrences of *keel* can be used to re-weight query results, ranking items based on co-occurrences. However, relationships among terms are not necessarily made explicit, although they can be inferred from the definition.

From Table 8.1, *sternpost* and *rabbet* are two terms that co-occur with keel. The ontology expresses several facts about these terms, such as: a) a keel can be rabetted, b) a rabbet is a technique, and c) a scarf can connect the keel to the sternpost. This information can be used to improve relevance by re-weighting queries depending on joining techniques and components the keel can be attached to. Figure 8.7 lists the rest of properties about keel defined in the ontology. A more complete depiction of the properties of keel and its relationships with other ship components is shown in Figure 8.8. The ontology lists the different shapes of a keel (square, vertical rectangle, T-shaped, rounded, tapered bottom surface, tapered or narrowing sides, and horizontal rectangle). A keel can be joined to a keel plank using a scarf, and a scarf in turn is a joining technique.

- hasAverageDoubleFrameMoldingNearMidBottom **Domain** **Keel**
- hasAverageDoubleFrameSidingAtKeel **Domain** **Keel**
- hasAverageFloorTimberMoldingNearMidBottom **Domain** **Keel**
- hasAverageFloorTimberSidingAtKeel **Domain** **Keel**
- hasChangedDimensionAft **Domain** (**Keel** **or** KeelPlank)
- hasChangedDimensionForward **Domain** (**Keel** **or** KeelPlank)
- hasMaximumMoldedDimension **Domain** (**Keel** **or** KeelPlank)
- hasMaximumSidedDimension **Domain** (**Keel** **or** KeelPlank)
- hasRabbetsCutIntoKeelSides **Domain** (**Keel** **or** KeelPlank)
- hasRabetsCutIntoUpperKeelEdges **Domain** (**Keel** **or** KeelPlank)
- hasSurvivalLength **Domain** **Keel**
- hasUnknownShaped **Domain** **Keel**
- isHorizontalRectangleShaped **Domain** **Keel**
- isKeelRockered **Domain** (**Keel** **or** KeelPlank)
- isRoundedOrTaperedBottomSurface **Domain** **Keel**
- isSquareShaped **Domain** **Keel**
- isTaperedOrNarrowingSides **Domain** **Keel**
- isTorVikingShaped **Domain** **Keel**
- isVerticalRectangleShaped **Domain** (**Keel** **or** KeelPlank)
- hasIntermediateScarf **Domain** (**Keel** **or** KeelPlank)
- hasKeelShape **Domain** **Keel**
- hasRabbetedTechnique **Domain** (**Keel** **or** KeelPlank)
- hasRockeredKeel **Domain** (**Keel** **or** KeelPlank)
- hasScarfConnectingStem **Domain** (**Keel** **or** KeelPlank)
- hasScarfConnectingSternpost **Domain** (**Keel** **or** KeelPlank)
- hasShoe **Domain** (**Keel** **or** KeelPlank)
- isKeelMadeOfWood **Domain** **Keel**

Figure 8.7. Properties of keel defined in *ontoShipDS*.

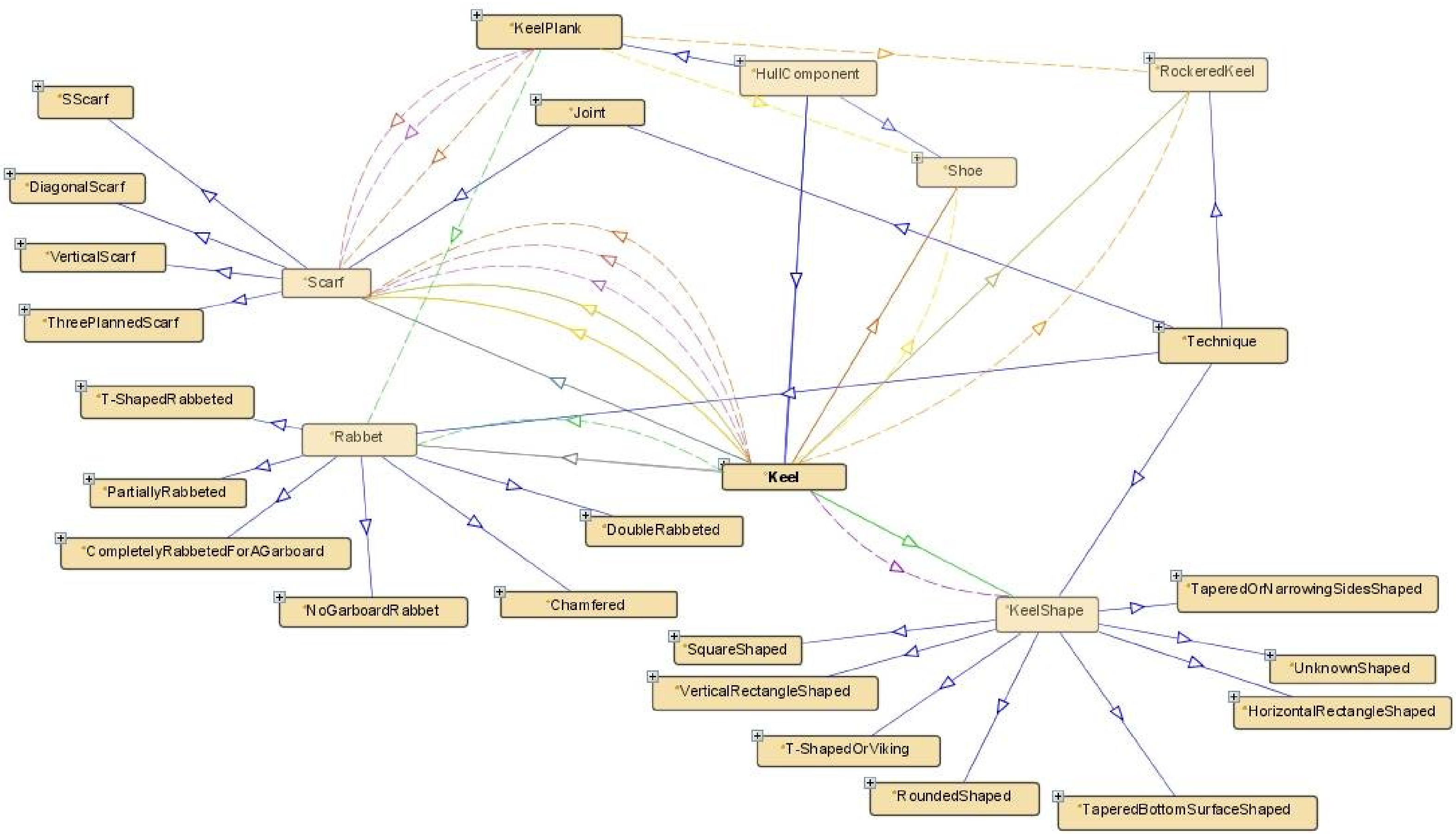


Figure 8.8. An extended graphic representation of some properties and relationships of keel defined in *ontoShipDS*.

*Plank i*s another case that highlights the limitations of the glossary and the full-text index. A search for *plank* and *planking* using the original Lucene search engine, returns the same results. Figure 8.9 shows a partial list of the results of searching for plank and planking (note that the total of hits is the same). Although this is the correct result from the functionality of a full-text index, it is not correct from the nautical archaeology standpoint. Plank is an object, while planking is a technique; therefore, they should be treated differently.

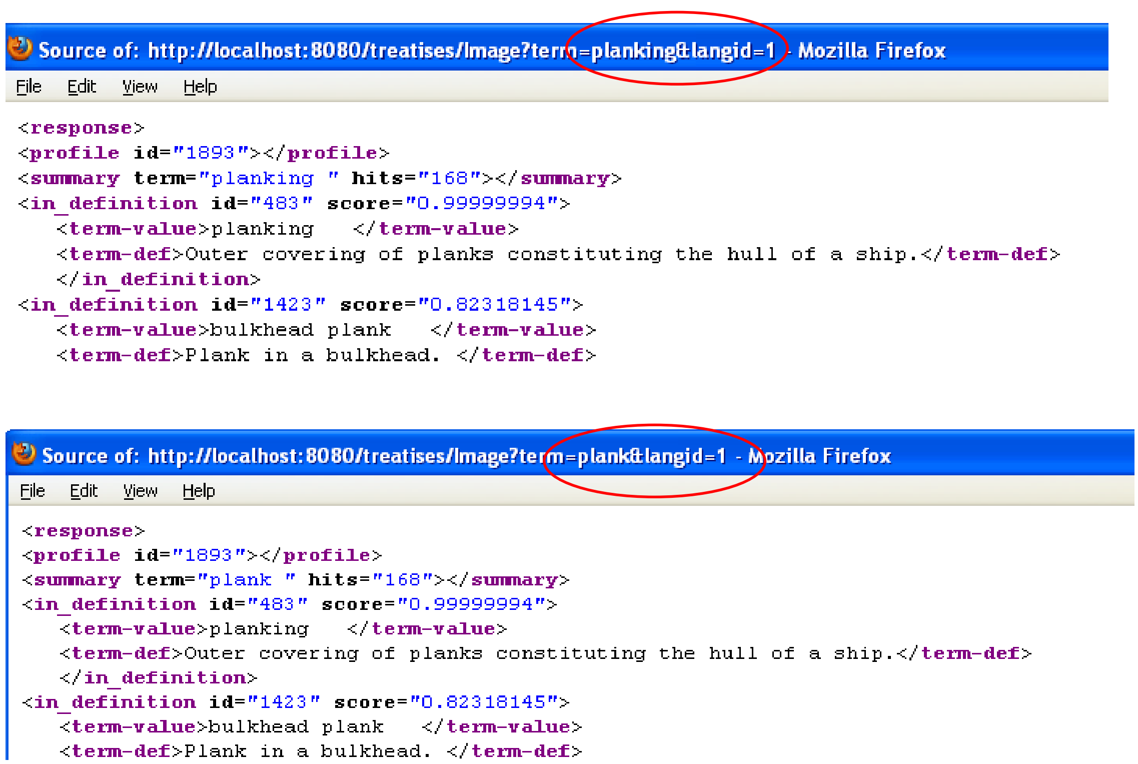


Figure 8.9. Results of the search for planking (top) and plank (bottom) generated by the expansion algorithm using Lucene.

In this case, the ontology can be used to ensure that a query search with the terms plank or planking is properly handled. Otherwise, we can at least inform the user about the different meanings of those terms. Figure 8.10 lists the properties of plank and planking as defined in the ontology.

- hasDiminishingStrakes **Domain Planking**
- isThereEvidenceOfRepair **Domain** (Frame **or Planking**)
- isThereEvidenceOfReplacement **Domain** (Frame **or Planking**)
- isTypeOfPlanking **Domain Planking**
- Plank ≡ (hasPlankRepairs **some** PlankingRepairs) **and** (hasPlankReplacements **some** PlankingReplacements) **and** (hasPlankToPlankJoinerySystem **some** PlankToPlankJoinerySystem) **and** (hasPlankWaleFasteningSystem **some** PlankWaleFastening) **and** (hasPlankingPattern **some** PlankingPattern) **and** (hasPlankingPlan **some** PlankingPlan) **and** (hasPlankingScarfs **some** Scarf) **and** (hasPlankingTechnique **some Planking**)
- isPlankingMadeOfWood **Domain Planking**

- hasPlankingPattern **Domain Plank**
- hasPlankingPlan **Domain Plank**
- hasPlankRepairs **Domain Plank**
- hasPlankReplacements **Domain Plank**
- hasPlankToPlankJoinerySystem **Domain Plank**
- hasPlankWaleFasteningSystem **Domain Plank**
- isFastened **Range Plank**

Figure 8.10. Properties of planking (top) and plank (bottom) as defined in the ontology.

Figure 8.11 depicts some of the relationships of plank and planking as defined in the ontology. It can be seen that *plank* is a hull component, while planking is a *technique*. Furthermore, the various planking techniques are listed, as well as other hull components. In the case of a plank, it states that it can be related to a strake and a wale (this relationship cannot be inferred in the glossary).

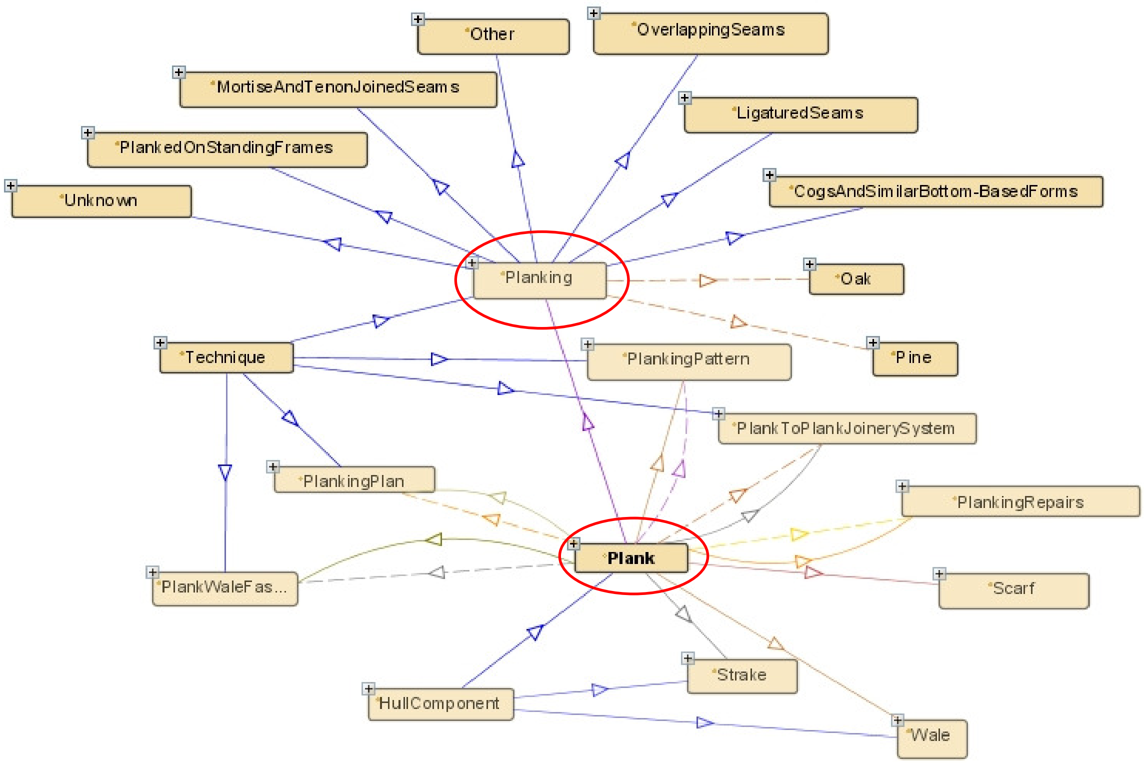


Figure 8.11. Graphical depiction of plank and planking as defined in the ontology.

8.7 Discussion

Some of the characteristics of shipbuilding treatises are their specialized domain, provenance (written in various languages), and time (spanning from the 16th to the 19th-century). These technical documents are used by nautical archaeologists to understand different shipbuilding techniques, and to compare their descriptions with archaeological evidence recovered from underwater excavations.

Although the number of transcriptions available to conduct the experiments was limited, it included written sources in various languages: Spanish, English, Portuguese, Italian, and Venetian. These sources offered a linguistically diverse collection for the evaluation.

Using additional information for query expansion showed improvements in the contextualization of terms. The original Lucene's ranking mechanism scores query results based on inverse document frequency. This approach however, is solely based on a combination of the number of times a term appears in a document and the number of documents in the collection where it appears. Therefore, semantic information is not considered. The first method (experiment 1) based on co-occurrence of terms in the corpus, although using term frequency; exploits semantic information. This is accomplished by considering terms in the glossary for scoring results.

Results from calculating co-occurrence showed higher frequency count and distance measure for Portuguese and Spanish texts than for English. This is the result of the characteristics of the English texts, which correspond to short narratives describing shipwrecks. Although, determining the degree in which this affects scoring, results from experiments seem to suggest that context window size (paragraphs in this case) impacts results.

Furthermore, some of the results in text co-occurrence were confirmed by information derived from the definitions of the terms in the glossary. Unlike corpus co-occurrence, information from definitions in the glossary (experiment 2) is a better source for determining semantic relationships among terms. This is based on the fact that a definition explicitly states what terms are related. Using the three definitions in experiment 1 (stern post, stempost, and floor timber) confirms this observation. The definition of stern post, indicates that timber, keel, planking, stern, and rudder are relevant terms, "A heavy *timber* on the *keel* end usually inclined slightly aft on which *planking* terminates and any *stern* frame is constructed, and on which the *rudder* is hung."

The other two definitions (stem post and floor timber) also show relevant terms for each one. Definition of stem post, for instance, suggests that timber, scarfed, keel, and plank are relevant, "a vertical or upward curving *timber* or assembly of timbers *scarfed* to the *keel* or central *plank* at its lower end, into which the two sides of the bow were joined." Conversely, frame, keel, and keelson are relevant terms to floor timber, "the central element of a *frame*. The *timber* that straddles the *keel* or *keelson*(s) upon which the remainder of the *frame* is build on." An important advantage of the glossary is the possibility of expanding related terms from definitions in different languages.

However, terms in the definitions do not indicate how they are related. This limitation is overcome by the ontology. Relationships and properties expressed in the ontology can be used for better contextualize terms, since the relationships are explicitly stated. For example, the ontology indicates that stern post is a hull component and a component (Figure 8.12). From the terms in its definition, the ontology indicates that planking is a technique (Figure 8.13), that there exist seven planking techniques, and three planking patterns. Properties of keel on the other hand are listed on Figure 8.7. The taxonomy of *rudder* shows it as an auxiliary component (Figure 8.14).

- + Thing
    - + Component
        - + HullComponent
            - **SternpostTimber**

Figure 8.12. Taxonomy of stern post according to *ontoShipDS*.

- + Thing
    - + Technique
        - + **Planking**
            - CogsAndSimilarBottom-BasedForms
            - LigaturedSeams
            - MortiseAndTenonJoinedSeams
            - Other
            - OverlappingSeams
            - PlankedOnStandingFrames
            - Unknown

- + Thing
    - + Technique
        - + **PlankingPattern**
            - DoublePlanked
            - SinglePlanked
            - TriplePlanked

Figure 8.13. Properties of planking according to *ontoShipDS*. Taxonomy of planking and planking techniques (top). Planking patterns (bottom).

Figure 8.14. *ontoShipDS* showing rudder.

Query results rankings vary depending on the weighting function used. This in turn enables the presentation of information in different contexts. Table 8.8 shows the scores for the three proposed methods for the term *keel*. The third column shows scores based on term co-occurrence in the texts (equation 8.1). The fourth and fifth columns show scores based on equations 8.2 and 8.3 respectively. Basing rankings on equation 8.2 indicates that hull, molded, and side will be ranked at the top (score 1.000); followed by floor timber, stern post, garboard, rabbet, bottom, heel, and sternpost (score 0.750); while strake, planking, top, mortise, and pine, will rank at the bottom (score 0.000). Terms that appear in the definition of keel receive the highest score (1.000), which indicates a strong relationship. Terms whose definitions include keel receive a score of 0.750, meaning that they are related but to a lesser degree. The remaining terms, the ones whose definition do not include keel will rank at the bottom (score 0.000).

Scores shown in the fifth column are based on equation 8.3, they indicate that stem post, garboard, strake, hull, molded, planking, mortise, pine, rabbet, and bottom will receive the highest score (1.000). These are terms for which the ontology has information relating them to keel. The rest of the terms: floor timber, top, and side (score 0.000) will rank at the bottom. These are terms for which the ontology lacks information that relates them to keel.

The ontology can also be used for ranking results based on other parameters. For example garboard, strake, sternpost, and heel will rank higher as *hull components*. Mortise and rabbet will rank higher as *joinery system*. Planking will rank higher as *technique*. Finally pine will rank higher as *wood property*. This characteristic differentiates the ontology from the glossary, and shows a much richer semantic description of nautical terms.

Table 8.8. Scoring results for keel.

Id	Term	method 1	method 2	method 3
56	floor timber	0.260	0.750	0.000
138	stem post	0.115	0.750	1.000
66	garboard	0.077	0.750	1.000
144	strake	0.067	0.000	1.000
79	hull	0.058	1.000	1.000
457	molded	0.058	1.000	1.000
483	planking	0.058	0.000	1.000
2503	top	0.058	0.000	0.000
546	side	0.048	1.000	0.000
460	mortise	0.038	0.000	1.000
1873	pine	0.038	0.000	1.000
114	rabbet	0.038	0.750	1.000
17	bottom	0.029	0.750	1.000
387	heel	0.029	0.750	1.000
143	sternpost	0.029	0.750	1.000

Table 8.9 depicts the terms sorted by descending score according to the three methods (these are the same scores from Table 8.8). Showing these results in the same table facilitates the visualization of how term ranks change according to the weighting

method used. Floor timber ranks 1, 4, and 12 for each one of the weighting methods. Method 1, although placing floor timber at the top, is probably not the strongest one from the semantic standpoint, since the weight is solely based on co-occurrence of terms. Method 2, on the other hand, has more semantic value, since it uses definitions from the glossary. Method 3, would be expected to give the highest score (because the ontology has the richest semantic description), instead scores 0.000. This apparent contradiction is because the ontology does not have any information about floor timber.

Table 8.9. Terms related to floor timber, sorted by score for each method.

id	term	method 1	id	term	method 2	Id	term	method 3
56	floor timber	0.260	79	hull	1.000	66	garboard	1.000
138	Stem post	0.115	457	molded	1.000	144	strake	1.000
66	garboard	0.077	546	side	1.000	79	hull	1.000
144	strake	0.067	56	floor timber	0.750	457	molded	1.000
2503	top	0.058	138	stem post	0.750	483	planking	1.000
483	planking	0.058	66	garboard	0.750	460Of	mortise	1.000
457	molded	0.058	114	rabbet	0.750	1873	pine	1.000
79	hull	0.058	17	bottom	0.750	114	rabbet	1.000
546	side	0.048	387	heel	0.750	17	bottom	1.000
1873	pine	0.038	143	sternpost	0.750	387	heel	1.000
460	mortise	0.038	144	strake	0.000	143	sternpost	1.000
114	rabbet	0.038	483	planking	0.000	56	floor timber	0.000
387	heel	0.029	2503	top	0.000	138	stem post	0.000
143	sternpost	0.029	460	mortise	0.000	2503	top	0.000
17	bottom	0.029	1873	pine	0.000	546	side	0.000

9. CONCLUSIONS

The use of information technology in general and digital libraries in particular has shown a great impact on numerous disciplines. What is probably not immediately visible are the challenges in making these collections available in electronic format, the tools required for transforming and aggregating the objects and artifacts that compose them, as well as the new scholarly practices created. In this context, this dissertation lies at the intersection of digital libraries and nautical archaeology.

The contribution of this dissertation is an approach for improving contextualization of nautical terms. This is accomplished by exploiting two external sources; namely, a specialized glossary in various languages, and an ontology of wooden ships. The methodology is based on deriving semantic information as a mechanism for relating terms in a more meaningful way. In addition, an algorithm for expanding queries in multiple languages is proposed.

The theoretical framework of this dissertation derives mainly from research in information retrieval and the semantic web; in particular, term co-occurrences, query expansion, and semantic information for improving query scoring. In this dissertation, two conceptualizations of ships were analyzed: textual and visual. Textual representations offer more detailed abstractions, while visual ones provide more contextual abstractions. The key from the standpoint of ship reconstruction (the chief problem addressed here), is in finding a way for combining both, and ultimately facilitating the understanding of a composite object (a vessel) from incomplete and damaged components.

From the perspective of text encoding, XML provided a flexible mechanism for structuring transcriptions and allowing the incorporation of external information. Therefore, definitions and translations from the glossary were embedded in the texts in order to facilitate their understanding. This is a useful feature given the nature of the domain (a technical specialized discipline), the use of ancient technical terms, and the diversity of languages.

The pioneering work of Dick Steffy highlights the importance of incorporating new technologies for the preservation of information in nautical archaeology. The workflow followed in the transformation of his original conceptualization into an RDF-OWL version, and included numerous revisions by domain experts in order to guarantee its contents.

In order to understand the effect of different approaches for improving contextualization, three methods are devised. These methods are based on text co-occurrence, glossary co-occurrence, and ontology-based. For each method, a weighting function was created that calculates scores of query results. Furthermore, an evaluation was conducted to analyze the impact of each method in ranking query results.

Evaluation results show that both co-occurrences from the texts and glossary improved the contextualization of information. However, the relationships were not explicitly stated. The ontology proved a good tool for making explicit knowledge about ships, in particular the relationships among terms in the domain. This in turn, enhanced the semantic information provided for re-weighting query results based on text and glossary co-occurrences.

Preliminary results from the experiments and properties of the supporting materials indicate that using co-occurrences and the ontology in combination rather than in isolation is a better approach because they complement each other. This also suggests that the glossary has certain properties that the ontology lacks, and the other way around. One property of the glossary is the ability to enable cross-lingual search, something the ontology does not provide. As expected, the ontology expresses more detailed semantic relationships. In addition, text co-occurrence results have a higher number of combinations that can be used for ranking results. However, there is no guarantee that text-based co-occurrences are technically related, although, some results were confirmed by the definitions. In contrast, co-occurrences from definitions in the glossary, while less numerous, explicitly indicate related terms.

Despite the results obtained in this study, there were some limitations. Certain languages in the glossary contained few entries. This results in one language being benefited over the others, as well in biased results in the use of co-occurrences. Similarly, term co-occurrence was based on frequency, although average distance could have also been chosen.

These limitations however, open the opportunity for future work. Experiments to evaluate the effect of using average distance instead of frequency, or a combined approach, could improve even further the contextualization of nautical terms.

Similarly, more experimentation and analysis are required for comparing the effects of using text co-occurrence in a particular language for ranking query results in other languages. The experiments in this study used co-occurrences in English because that was the language with most numerous entries in the glossary.

The ontology showed richer descriptions of components in the domain. However, being limited to one language made it difficult to be used in query expansion in other languages. This shortcoming was overcome with the use of the glossary.

Working in a real domain brought interesting results. The architecture of *glosShip* was created with scalability in mind. It was rewarding to see the architecture scaling well when Venetian was added to our existing languages. However, an expert editor encountered problems when translating words. For example, he mentioned the richness in expressing medieval anchors, and the impossibility of matching terms one-to-one across various languages. This in turn opens an interesting research area from the computational linguistics standpoint.

As stated earlier, the goal in this dissertation is a preliminary step in using additional information in the contextualization of terms in a domain. In future research, it would be useful to conduct user studies, in order to evaluate real scenarios.

The creation of a mechanism for using textual and visual abstractions simultaneously and measuring their effect is another area to explore, especially because it can be extended to collections where visual and textual representations of the same object can be combined.

Using additional sources for computing term co-occurrence can help to better understand how particular materials in the collection influence this calculation. I foresee in the near future the incorporation of some thesis and dissertations in nautical archaeology to the collection, which in turn adds a new genre of texts.

With the existing tagged images (both texts and illustrations), the glossary and the ontology can be used for automatic tagging. Labels from this process will not be limited to terms; they can include properties and relationships from the ontology. This could be a preliminary step in the creation of illustrated dictionaries in any language.

REFERENCES

ACEVEDO, D., VOTE, E., LAIDLAW, D., AND JOUKOWSKY, M. 2000. ARCHAVE: A virtual environment for archaeological research. In *Proceedings of the 28th Annual Internatinal Conference of Computer Applications.* 18-21.

ACEVEDO, D., VOTE, E., LAIDLAW, D., AND JOUKOWSKY, M. 2001. Archaeological data visualization in VR: Analysis of lamp finds at the Great Temple of Petra, a case study. In *Proceedings of the Conference on Visualization.* 493-496.

ADRIANI, M., AND VAN RIJSBERGEN, C. 1999. Term similarity-based query expansion for cross-language information retrieval. *Lecture Notes in Computer Science*. *1696*, 311-322.

THE APACHE LUCENE PROJECT. 2009.
http://lucene.apache.org/ (Accessed on 11/09).

APPELT, W. 1988. FODA—The formal specification of ODA document structures and its use as a basis for conformance testing. *Computer Standards & Interfaces. 7*, 4, 377-385.

APPELT, W., CARR, R., AND TETTEH-LARTEY, N. 1990. Conformance testing of ODA socuments and ODA implementations according to ISO 8613 based on its formal description (FODA). *Computer Standards & Interfaces. 11*, 1, 47-55.

ASHBURNER, M., BALL, C., BLAKE, J., BOTSTEIN, D., BUTLER, H., CHERRY, J., DAVIS, A., DOLINSKI, K., DWIGHT, S., EPPIG, J., HARRIS, M., HILL, D., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J., RICHARDSON, J., RINGWALD, M., RUBIN, G., AND SHERLOCK, G. 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics. 25*, 25-29.

ASHMORE, W., AND SHARER, R. 2003. *Discovering Our Past: A Brief Introduction to Archaeology*. McGraw-Hill, New York.

BAADER, F., AND NUTT, W. 2003. Basic Description Logics. In *The Description Logic Handbook: Theory, Implementation and Applications.* Baader, F., Calvanese, D., McGuiness, D., Nardi, D., and Patel-Schneider, P. (Eds.) Cambridge University Press. Cambridge, UK.

BACA, M. 2004. Fear of authority? Authority control and thesaurus building for art and material. *Cataloguing & Classification Quarterly. 38*, 3-4, 143-151.

BASCH, L. 1972. Ancient wrecks and the archaeology of ships. *International Journal of Nautical Archaeology. 2*, 1, 137-145.

BETHUNE, J. 1983. *Technical Illustration*. John Wiley & Sons, New York.

BLOMERUS, P., AND LESK, A. 2008. Using AutoCAD to construct a 4D block-by-block model of the Erechtheion on the Akropolis at Athens, II: Connecting a database to an AutoCAD Model. *CSA Newsletter. 20*, 3.

BROWN, M., AND SEALES, W. 2000. Beyond 2D images: Effective 3D imaging for library materials. In *Proceedings of the 5th ACM Conference on Digital Libraries.* 27-36.

BROWN, M., AND SEALES, W. 2001. The digital Atheneum: New approaches for preserving, restoring and analyzing damaged manuscripts. In *Proceedings of the 1st Joint Conference on Digital Libraries*. 437-443.

BRUSA, G., CALIUSCO, M., AND CHIOTTI, O. 2006. A process for building a domain ontology: An experience in developing a government budgetary ontology. In *Proceedings of the Australasian Ontology Workshop.* 7-15.

THE CANTERBURY TALES PROJECT. 2008. http://www.cta.dmu.ac.uk/projects/ctp/index.html (Accessed on 6/08).

CARBON, S., IRELAND, A., MUNGALL, C. J., SHU, S., MARSHALL, B., AND LEWIS, S. 2008. AmiGO: Online access to ontology and annotation data. *Bioinformatics. 25*, 2, 288-289.

CASTRO, F. 1998. The remains of a portuguese indiaman at Tagus Mouth, Lisbon, Portugal (*Nossa Senhora dos Mártires*, 1606?). In *Proceedings of the International Symposium: Archaeology of Medieval and Modern Ships of Iberian-Atlantic Tradition.* 381-404.

CASTRO, F. 2003. The *Pepper Wreck*. *International Journal of Nautical Archaeology. 32*, 1, 6-23.

CASTRO, F. 2005a. *The Pepper Wreck*. Texas A&M University Press, College Station, TX.

CASTRO, F. 2005b. Rigging the *Pepper Wreck*. Part I: Masts and yards. *International Journal of Nautical Archaeology*. *34*, 1, 112-124.

CASTRO, F. 2005c. The Pepper Wreck: Nossa Senhora dos Martires, Lisbon, Portugal. In *Beneath the Seven Seas*. Bass, G. (Ed.) Thames & Hudson. London, UK.

CHIMAERA. 2000. Chimaera Ontology Environment www.ksl.stanford.edu/software/chimaera (Accessed on 2/09).

CIULA, A., SPENCE, P., AND VIEIRA, J. 2008. Expressing complex associations in medieval historical documents: The Henry III fine rolls project. *Literary and Linguistic Computing. 23*, 3, 311-325.

COFFEY, P. 1914. *Ontology or the Theory of Being: An Introduction to General Metaphysics*. Longmans, London, UK.

COMMERCIAL SHIP ONTOLOGY. 2005. http://orlando.drc.com/semanticweb/daml/ontology/commercial/shipping/commercialship-ont (Accessed on 2/09).

CRANE, G. 1988. The Perseus project: An interactive curriculum on classical Greek civilization. *Educational Technology. 28*, 11, 25-32.

CRANE, G. 1996. Building a digital library: The Perseus project as a case study in the humanities. In *Proceedings of the 1st ACM International Conference on Digital Libraries*. 3-10.

CRANE, G. 2002. Cultural heritage digital libraries: Needs and components. In *Proceedings of the European Conference on Digital Libraries.* 626-637.

CUTLER M., DENG H., MANICCOM S., AND MENG W. 1999. A new study on using HTML structure to improve retrieval. In *Proceedings of the 11th IEEE Conference on Tools with Artificial Intelligence.* 406-409.

DAY-RICHTER, J., HARRIS, M. A., AND HAENDEL, M. 2007. OBO-Edit—an ontology editor for biologists. *Bioinformatics. 23*, 16, 2198-2200.

DEROSE, S., DURAND, D., MYLONAS, E., AND RENEAR, A. 1990. What is text, really? *Journal of Computing in Higher Education. 1*, 2, 2-36.

THE DIGITAL IMPRINT. 2008. Institute of Archeology. University of California at Los Angeles. http://www.sscnet.ucla.edu/ioa/labs/digital/imprint/proposal.html (Accessed on 6/08).

DOERR, M., AND CROFTS, N. 1999. Electronic esperanto: The role of the object oriented CIDOC reference model. In *Proceedings of the 5th International Cultural Heritage Informatics Meeting*. 157-173.

EGAN, V., BIKAI, P., AND ZAMORA, K. 2000. Archaeology in Jordan. *American Journal of Archaeology. 104*, 3, 561-588.

EIRON, N. AND MCCULEY, K. 2003. Analysis of anchor text for web search. In *Proceedings of the 26th Conference on Information Retrieval.* 459-460.

FOERTSCH, R. 2006. ARACHNE - datenbank und kulturelle archive des forschungsarchivs fuer antike plastik koeln und des deutschen archaeologischen instituts. http://arachne.uni-koeln.de/ (Accessed on 8/09).

FORTE, M. 1997. Introduction. In *Virtual archaeology: Re-creating ancient worlds*. Forte, M., and Siliotti, A., (Eds.) Harry N. Abrams. New York.

FRAGA, T. 2007. Santo Antonio de Tanna: Story and reconstruction. M.A. thesis, Texas A&M University, College Station, TX.

FU, G., JONES, C., AND ABDELMOTY, A. 2005. Ontology-based spatial query expansion in information retrieval. *Lecture Notes in Computer Science*. *3761*, 1466-1482.

GAZENDAM, L., MALAISÉ, V., SCHREIBER, G., AND BRUGMAN, H. 2006. Deriving semantic annotations of an audiovisual program from contextual texts. In *Proceedings of 1st International Workshop on Semantic Web Annotations for Multimedia*. http://www.cs.vu.nl/~guus/papers/Gazendam06a.pdf (Accessed on 7/09).

GOGGIN, M. 2004. *Defining Visual Rhetorics*. Lawrence Eslbaum Associates, Mahwah, NJ.

GREEN J., AND SOUTER, C. 2002. Archaeological application of the HPASS (high precision acoustic surveying system) to surveys of the HMS pandora wreck-site and the roman bridge at Maastricht. *International Journal of Nautical Archaeology. 31*, 2, 273-282.

GREEN J., MATTHEWS, S., AND TURANLI, T. 2002. Underwater archaeological surveying using PhotoModeler, virtual mapper: Different applications for different problems. *International Journal of Nautical Archaeology. 31*, 2, 283-292.

GREG, W. 1933. The function of bibliography in literary criticism illustrated in a study of the text of "King Lear." *Neophilologus. 18*, 1, 241-262.

GRUBER, T. 1993. Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies. 43*, 5-6, 907-928.

GRÜNINGER, M., AND FOX, M. 1995. Methodology for the design and evaluation of ontologies. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*. 1-10.

GUARINO N., AND GIARETTA, P. 1995. Ontologies and knowledge bases: Towards a terminological clarification. In *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*. Mars, N. (Ed.) IOS Press. Amsterdam, The Netherlands.

HADINGHAM, E. 2000. Secrets of a desert metropolis: The hidden wonders of Petra's ancient engineers. *Scientific American Discovering Archaeology. 2*, 4, 70-74.

HARMAN, D. 1992. Relevance feedback revisited. In *Proceedings of the 15th International Conference on Information Retrieval.* 1-10.

HAZLETT, A. 2006. The nau of the livro nautico: The textual excavation of a Portuguese indiaman. In *The Edge of Empire*. Castro, F., and Custer, K. (Eds.) Caleidoscopio. Portugal.

HAZLETT, A. 2007. The Nau of the Livro Náutico: Reconstructing a sixteenth-century indiaman from texts. Ph.D. diss., Texas A&M University.

HOCKEY, S. 1999. Panel with Allen Renear and Jerome McGann: 'What is Text? A debate on the philosophical and epistemological nature of text in the light of humanities computing research. http://www.humanities.ualberta.ca/Susan_Hockey/achallc99.htm (Accessed on 11/08).

HOLT, P. 2008. *Site Recorder Database Schema*. 3H Consulting Ltd. Software. Plymouth, UK.

HUITFELDT, C. 1995. Multi-dimensional texts in a one-dimensional medium. *Computers and the Humanities. 28*, 4-5, 235-241.

JOUKOWSKY, M. 1993. Archaeological excavations and survey of the southern temple at Petra, Jordan. *Annual of the Department of Antiquities of Jordan. 38*, 293-322.

KAWAZOE, A., CHANLEKHA, H., SHIGEMATSU, M., AND COLLIER, N. 2008. Structuring an event ontology for disease outbreak detection. *BMC Bioinformatics. 9*, 3, S8.

KENSEK, K., SWARTZ, L., AND CIPOLLA, N. 2004. Fantastic reconstruction or reconstructions of the fantastic? Tracking and presenting ambiguity, alternatives, and documentation in virtual worlds. *Automation in Construction. 13*, 175-186.

KIRSH, D. 2002. Why illustrations aid understanding. In *Proceedings of the International Workshop on Dynamic Visualizations and Learning*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.5322&rep=rep1&type=pdf (Accessed on 7/09).

KRAFT, H. 1990. *Editionsphilologie*. Wissentschaftliche buchgeselschaf. Darmstadt, Germany.

LAME, G., AND DESPRES, S. 2005. Updating ontologies in the legal domain. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law.* 155-162.

LEE, K. AND LEE, S. 2000. An agent-based approach to preliminary ship design. *Journal of Marine Science and Technology. 5*, 2, 78-88.

LESK, A. 2004. A diachronic examination of the erechtheion and its reception. Ph.D. diss., University of Cincinnati, OH.

MACHADO, J., MARTINS, B., AND BORBINHA, J. 2009. Experiments on a multinomial language model versus Lucene's off-the-shelf ranking scheme and Rocchio query expansion (TEL@CLEF Monolingual Task). *Lecture Notes in Computer Science. 5706*, 50–57.

MISSIKOFF, M., VELARDI, P., AND FABRIANI, P. 2003. Text mining techniques to automatically enrich a domain ontology. *Applied Intelligence. 18*, 3, 323-340.

MOLYNEAUX, B. 1997. Introduction. In *The Cultural Life of Images: Visual Representation in Archaeology*. Molyneaux, B., (Ed.) Routledge. London, UK.

MONROY, C., FURUTA, R., AND CASTRO F. 2007a. A multilingual approach to technical manuscripts: 16th and 17th-century Portuguese shipbuilding treatises. In *Proceedings of the 7th Joint Conference on Digital Libraries.* 413-414.

MONROY, C., FURUTA, R., AND CASTRO F. 2007b. Texts, illustrations, and physical objects: The case of ancient shipbuilding treatises. *Lecture Notes in Computer Science. 4675*, 198-209.

MONROY, C., FURUTA, R., AND STRINGER, G. 2007. Digital Donne: Workflow, editing tools, and the reader's interface of a collection of 17th-century English poetry. In *Proceedings of the 7th Joint Conference on Digital Libraries.* 411-412.

NAVIGLI, R., AND VELARDI, P. 2003. An analysis of ontology-based query expansion strategies. In *Proceedings of the 14th European Conference on Machine Learning*. http://www.dcs.shef.ac.uk/~fabio/ATEM03/navigli-ecml03-atem.pdf (Accessed on 7/09).

NOY, N. AND MCGUINESS, D. 2001. Ontology development 101: A guide to creating your first ontology. Tech. rep. SMI-2001-0880, Stanford Medical Informatics. 1-25.

ONTOLINGUA. 1997. Ontolingua Reference Manual. http://www-kslsvc.stanford.edu:5915/doc/frame-editor/index.html (Accessed on 2/09).

ONTOLOGY BROWSER. 2009. http://code.google.com/p/ontology-browser/ (Accessed on 3/09).

PÉREZ-AGÜERA, J., AND ZARAGOZA, H. 2008. UCM-Y!R at CLEF 2008 robust and WSD tasks. In *Proceedings of the CLEF 2008 Workshop.* 137-143.

PROTÉGÉ. 2000. The Protege Project. http://protege.stanford.edu (Accessed on 1/10).

REEVES, C. 1992. *After Tut'ankhamun: Research and Excavation in the Royal Necropolis at Thebes*. Kegan Paul International, London, UK.

REEVES, N. AND WILKINSON, R. 1996. *The Complete Valley of the Kings: Tombs and Treasures of Egypt's Greatest Pharaohs*. Thames and Hudson, London, UK.

RENFREW, C. AND BAHN, P. 2000. *Archaeology. Theories, Methods and Practice*. Thames and Hudson, London, UK.

RHINOCEROS. 2009. Modeling Tools for Designers http://www.rhino3d.com/ (Accessed on 11/08).

RIETH, E. 1984. Principe de construction 'charpente premiere' et procedes de construction 'borde premier' au XVIIe siecle. *Neptunia. 153*, 21-31.

RIETH, E. 1998. Construction Navale à Franc-Bord en Méditerranée et Atlantique (XIVe-XVIIe siècle) et Signatures Archictecturales. Une Première Approche Archéologique. Méditerranée Antique. In *Pêche, Navigation, Commerce*. Rieth, E. (Ed.) CTHS. Paris, France.

ROCKWELL, G. 2003. What is text analysis, really? *Literary and Linguistic Computing. 18*, 2, 209-219.

THE ROSSETTI ARCHIVE. 2006. http://www.iath.virginia.edu/rossetti/ (Accessed on 11/06).

SCHREIBER, G., WIELINGA, B., AND JANSWEIJER, W. 1995. The KACTUS view on the 'O' word. In *Proceedings of International Joint Conference on Artificial Intelligence*. 159-168.

SCHWEIGHOFER, E., AND GEIST, A. 2007. Legal query expansion using ontologies and relevance feedback. In *Proceedings of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques. 321*, 149-160.

SHAPE. 2008. The SHAPE Lab @ Brown University. http://www.lems.brown.edu/vision/extra/shape/ (Accessed on 6/08).

SIMONETTA, A. 1998. *Nossa Senhora dos Mártires: The Last Voyage.* Verbo, Lisbon, Portugal.

SMITH, B. 2003. Ontology. In *The Blackwell Guide to Philosophy of Computing and Information.* Floridi, L. (Ed.) Wiley-Blackwell. Malden, MA.

SNYDER, L. 2004. Real-time visual simulation models in an exhibition environment. http://resumbrae.com/vr04/snyder.pdf (Accessed on 7/09).

SNYDER, L. 2008. The devil in the details: Reconstructing Chicago's white city. 2008 Colloquium on Digital Humanities and Computer Science. http://lucian.uchicago.edu/blogs/dhcs2008/files/2008/09/snyder.pdf (Accessed on 7/09).

SOERGEL, D. 1999. The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*. *50*, 12, 1119-1120.

SOO, V., CHEN-YU, L., CHAO-CHUN, Y., AND CHING-CHIH, C. 2002. Using sharable ontology to retrieve historical images. In *Proceedings of the ACM/IEEE International Conference on Digital Libraries.* 197-198.

STEFFY, R. 1994. *Wooden Ship Building and the Interpretation of Shipwrecks*. Texas A&M University Press, College Station, TX.

STEFFY, R. 2003. *Ship Data Project: A Compendium of Wooden Ship and Boat Construction Information Derived from Archaeological Excavations and Abandoned Hull Remains.* Unpublished raw data. The Institute of Nautical Archaeology. Texas A&M University, College Station, TX.

STUDER, R., BENJAMINS, V., AND FENSEL, D. 1998. Knowledge engineering: Principles and methods. *Data and Knowledge Engineering. 25*, 161-197.

THEBAN MAPPING PROJECT. 2008. American University in Cairo. http://www.thebanmappingproject.com/about/ (Accessed on 11/08).

THOMAS, D. H. 1998. *Archaeology*. Harcourt Brace College Publishers, Fort Worth, TX.

TRINKUNAS, L., AND VASILECAS, O. 2007. Building ontologies from relational databases using reverse engineering methods. In *Proceedings of the International Conference on Computer Systems and Technologies. 2*, 6, 1-6.

TUDORACHE, T. 2006. Employing ontologies for an improved development process in collaborative engeneering. Ph.D. diss., Technischen Universitat Berlin.

TUFTE, E. 2001. *Envisioning Information*. Graphic Press, Chesire, CT.

US NAVY SHIPS ONTOLOGY. 2002. http://www.daml.org/2002/03/usnships/ship-ont (Accessed on 11/09).

VAN DUIVENVOORDE, W. 2009. More than just bits of hull: Expensive oak, laminate construction, and goat hair. New insights on Batavia's archaeological hull remains. *Tijdschrift voor Zeegeschiedenis. 28*, 2, 59-68, 72-73.

VECHTOMOVA, O. AND KARAMUFTUOGLUB, M. 2007. Query expansion with terms selected using lexical cohesion analysis of documents. *Information Processing & Management. 43*, 4, 849-865.

VECHTOMOVA O. AND WANG Y. 2006. A study of the effect of term proximity on query expansion. *Journal of Information Science. 32*, 4, 324-333.

WELLS, A. 2008. Virtual reconstruction of a seventeenth century Portuguese nau. M.S. thesis, Texas A&M University, College Station, TX.

WIESMAN, F., ROOS, N., AND VOGT, P. 2002. Automatic ontology mapping for agent communication. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems.* 563-564.

XU, J. AND CROFT, B. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval.* 4–11.

ZUNIGA, G. 2001. Ontology: Its transformation from philosophy to information systems. In *Proceedings of the International Conference on Formal Ontology in Information Systems.* 187-197.

APPENDIX A

IMAGES OF *ontoShipDS*

This Appendix contains figures depicting various concepts and components in *ontoShipDS*. Each figure is preceded by a brief description. Unless stated otherwise images were created using Jumbalaya or OWL Viz, two ontology visualization plug-ins used with Protégé.

Depicting instances that contain the word Nail. It shows how Nails are used in various fastenings and joinery systems.

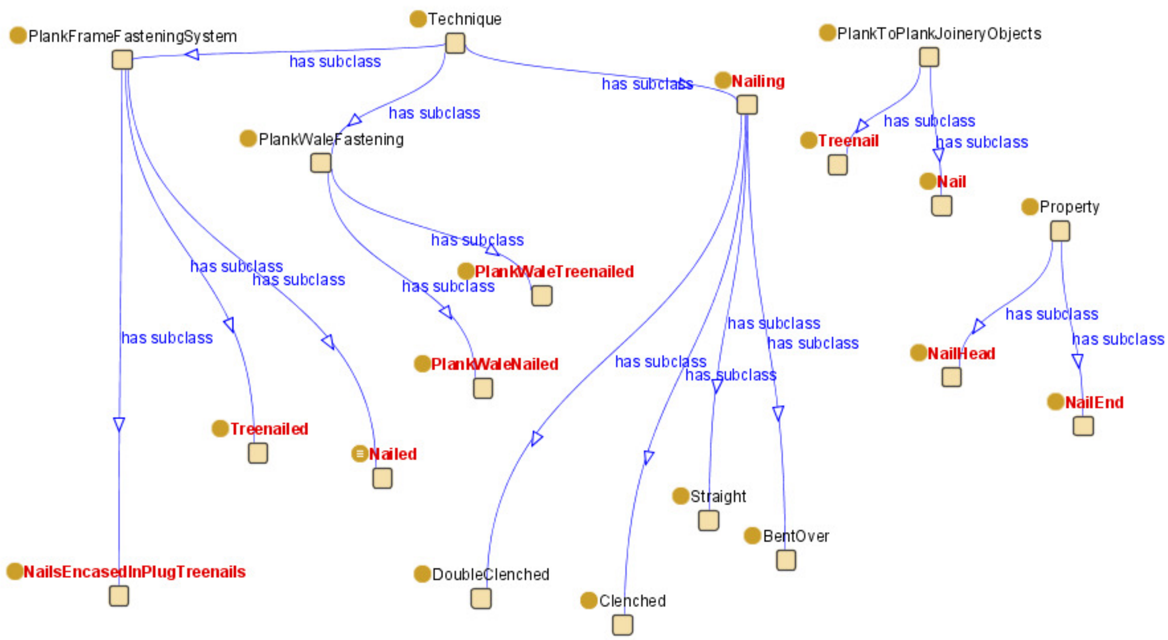
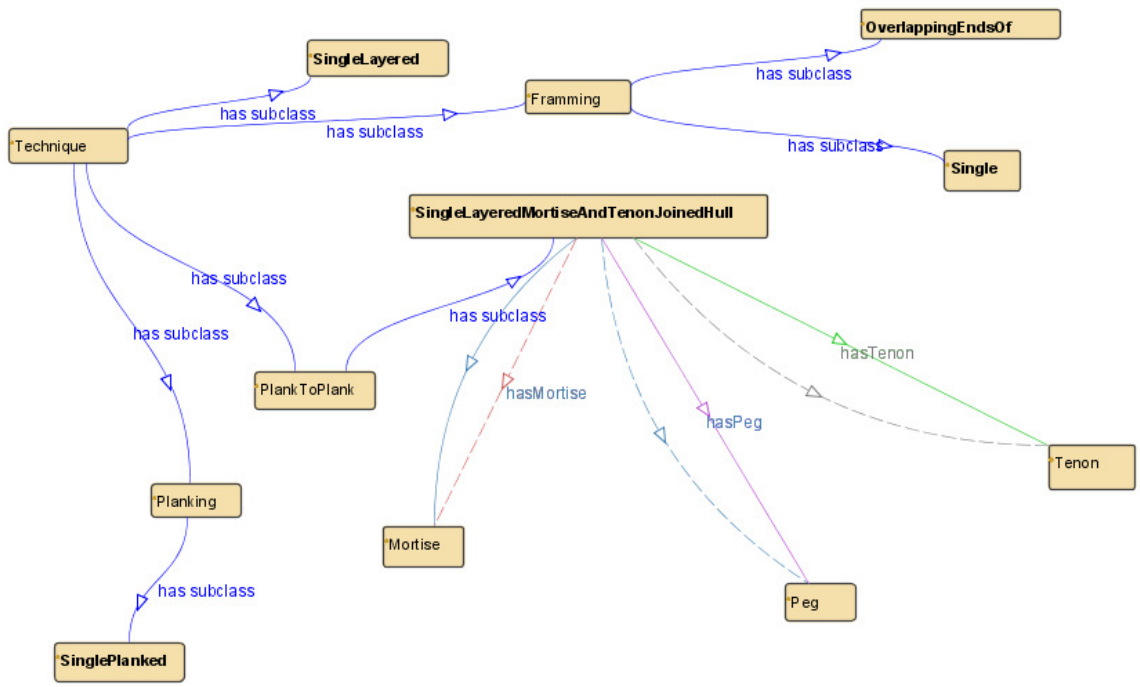
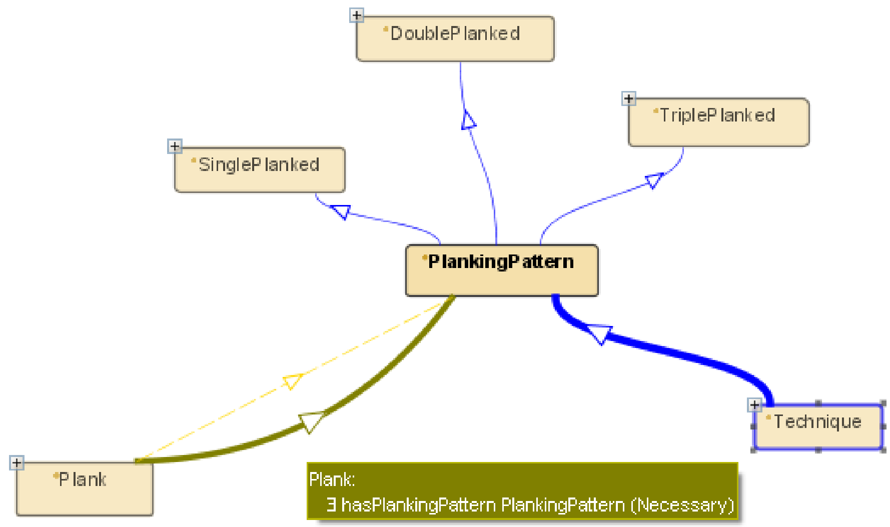


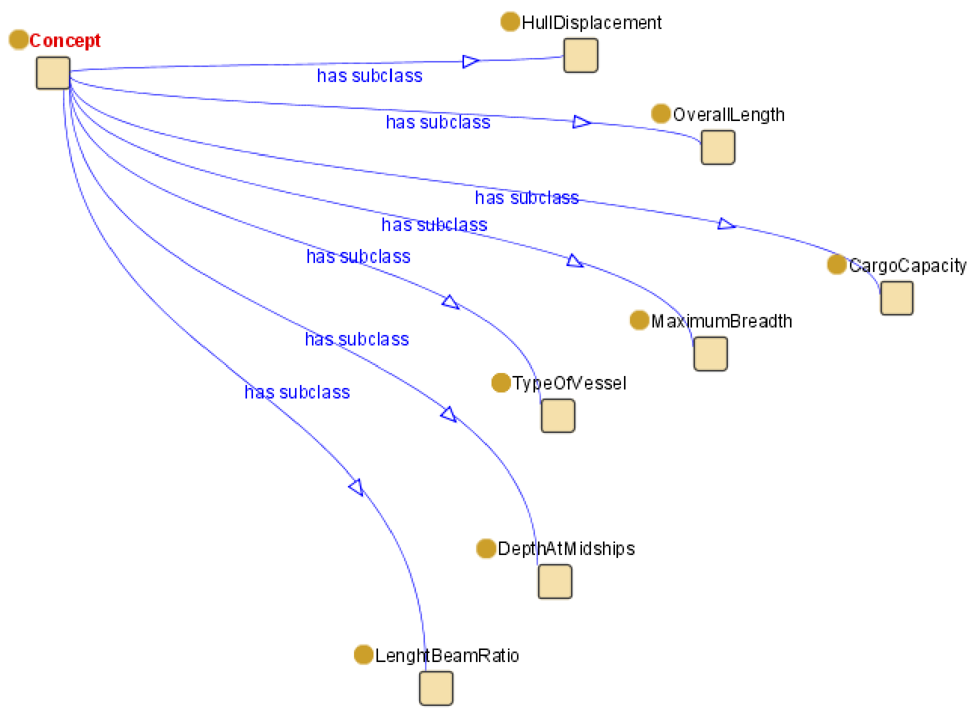
Figure that illustrates some techniques used in assembling planks with planks. For instance, a single layered hull can be joined using mortise, tenon, and pegs.



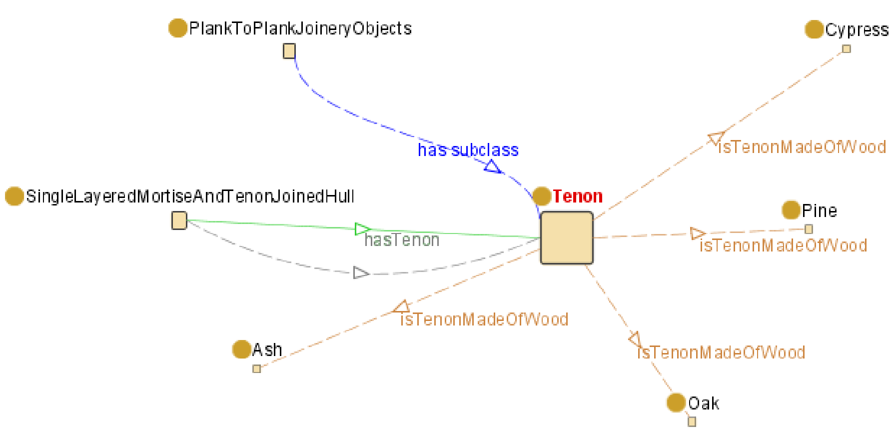
Depicting various planking patterns: single, double, and triple. The condition in green states that a Plank can be used in any of the planking patterns previously mentioned.



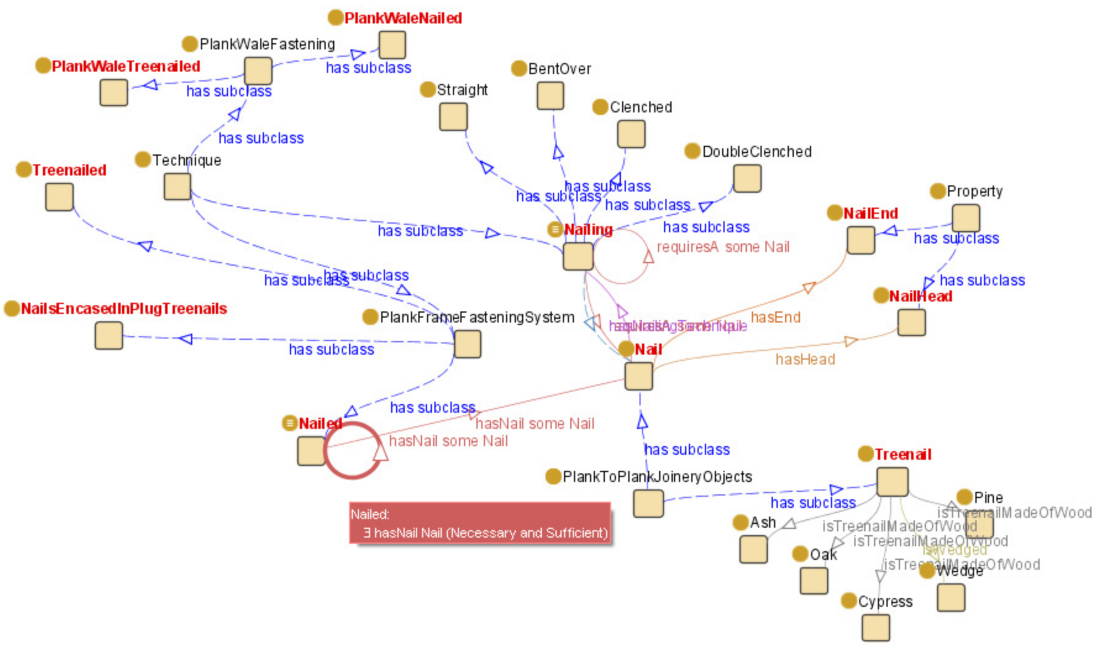
A graph depicting various concepts pertaining to ships as a whole.



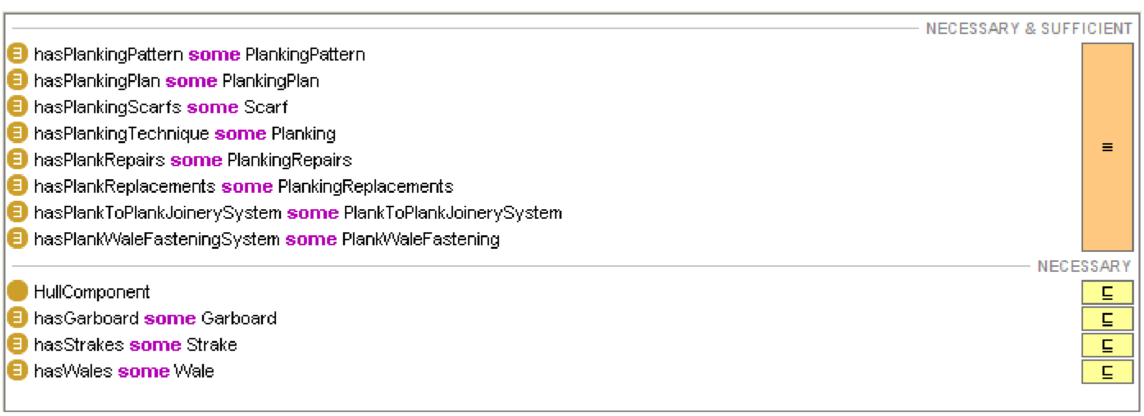
An image illustrating properties and relationships of Tenon. For example Tenons are used to join planks with planks. A single layered mortise and tenon joined hull requires a tenon. Also that it is a wooden object that can be made of cypress, pine, oak, or ash.



Depicting instances of nail, nailed, and nailing. It can be said that a nail can be used to join plank with plank. That planks with wales can be nailed or treenailed. The technique nailed that fastens plank and frame requires a nail.



An illustration listing the conditions that define Plank.



Listing properties of plank. The property at the bottom states that a plank plays a structural role in the vessel.

- hasGarboard (someValuesFrom Garboard)
  - Garboard
- hasPlankingPattern (single PlankingPattern) (someValuesFrom PlankingPattern)
  - PlankingPattern
- hasPlankingPlan (single PlankingPlan) (someValuesFrom PlankingPlan)
- hasPlankingScarfs (someValuesFrom Scarf)
- hasPlankingTechnique (someValuesFrom Planking)
  - Planking
- hasPlankRepairs (single PlankingRepairs) (someValuesFrom PlankingRepairs)
- hasPlankReplacements (single PlankingReplacements) (someValuesFrom PlankingReplacements)
- hasPlankToPlankJoinerySystem (single PlankToPlankJoinerySystem) (someValuesFrom PlankToPlankJoinerySystem)
- hasPlankWaleFasteningSystem (single PlankWaleFastening) (someValuesFrom PlankWaleFastening)
- hasStrakes (someValuesFrom Strake)
- hasWales (someValuesFrom Wale)
- playsAStructuralRole

The definition of Plank using the ontology browser.

Class: Plank

http://nadl.tamu.edu/nadlOntology/ship-ontology-1.0#Plank

Asserted Class Hierarchy

+ Thing
  + Component
    + HullComponent
      - **Plank**

Equivalents (1)

- (hasPlankRepairs **some** PlankingRepairs) **and** (hasPlankReplacements **some** PlankingReplacements) **and** (hasPlankToPlankJoinerySystem **some** PlankToPlankJoinerySystem) **and** (hasPlankWaleFasteningSystem **some** PlankWaleFastening) **and** (hasPlankingPattern **some** PlankingPattern) **and** (hasPlankingPlan **some** PlankingPlan) **and** (hasPlankingScarfs **some** Scarf) **and** (hasPlankingTechnique **some** Planking)

Superclasses (4)

- HullComponent
- hasGarboard **some** Garboard
- hasStrakes **some** Strake
- hasWales **some** Wale

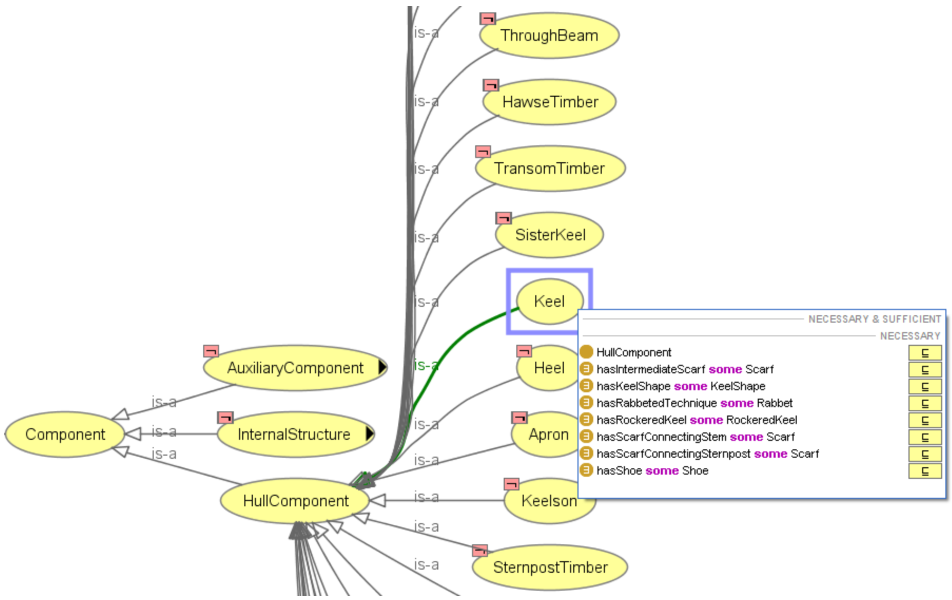
Disjoints (30)

Apron, Beam, BilgeKeel, Carling, Ceiling, Clamp, Deadwood, Decking, DiminishingStrake, Frame, Garboard, HawseTimber, Heel, Keel, KeelPlank, Keelson, Knee, RiderFrame, RiderKeel, RiderKeelson, Shoe, SisterKeel, StemTimber, SternpostTimber, Strake, Stringer, ThroughBeam, Transom, TransomTimber, Wale

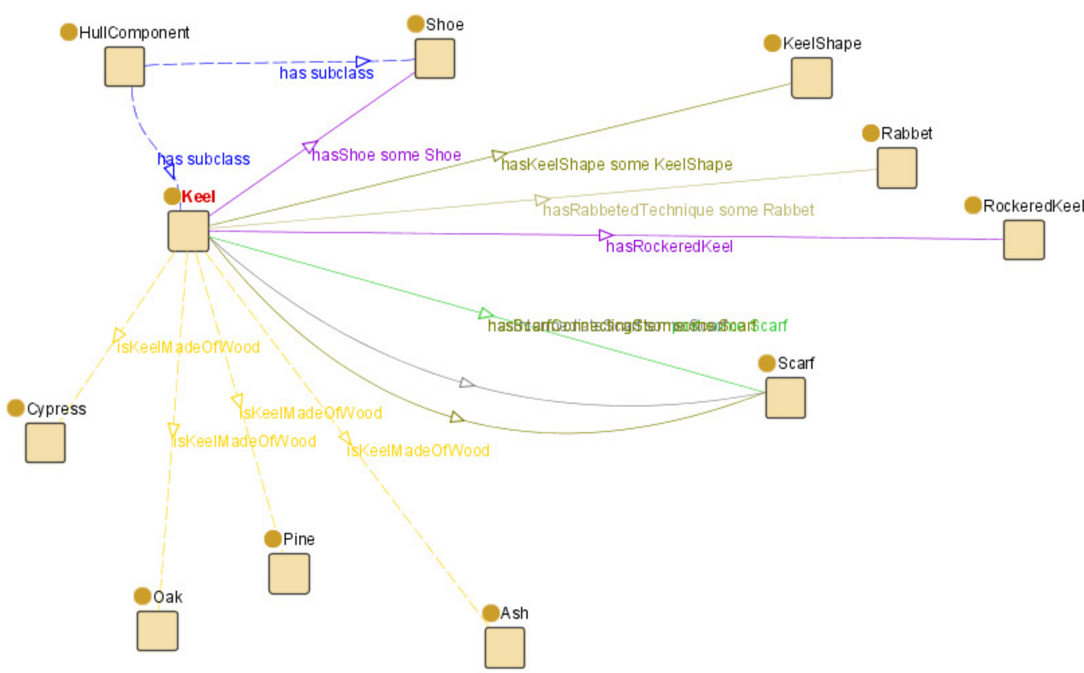
Usage (7)

- hasPlankingPattern **Domain** **Plank**
- hasPlankingPlan **Domain** **Plank**
- hasPlankRepairs **Domain** **Plank**
- hasPlankReplacements **Domain** **Plank**
- hasPlankToPlankJoinerySystem **Domain** **Plank**
- hasPlankWaleFasteningSystem **Domain** **Plank**
- isFastened **Range** **Plank**

Depicting information about Keel. It can be inferred that the keel is a hull component, and it can be assembled to the Stern and Stem post using Scarfs.



An extended version of the properties and relationships of the keel. Three properties indicate that, the keel is a wooden component, that a keel is a hull component, and that it can have a shoe.



APPENDIX B

SHIPBUILDING TREATISES CONSULTED AND *glosShip*

List of shipbuilding treatises used in this dissertation.

Title	Author	Provenance	Language	Published
The Complete Ship-wright	Edmund Bushnell	England	English	1678
Des Bois Propres au Service des Arsenaux	P. E. Herbin de Halle	France	French	1813
L'Architecture Navale	F. Dassie	France	French	1677
Traite´ d'Architecture Navale	Joseph Furttenbach	France	French	1629
Architectura Navalis	Joseph Furttenbach	Germany	German	1629
Nautica Mediterranea	Bartolomeo Crescenzio Romano	Italy	Italian	1601
Advertências de Navegantes	Marcos Cerveira de Aguilar	Portugal	Portuguese	1640
Coriosodades	Gonçallo de Sousa	Portugal	Portuguese	c. 1630
História Trágico-Marítima	Bernardo Gomes de Brito	Portugal	Portuguese	1552
História Trágico-Marítima	Bernardo Gomes de Brito	Portugal	Portuguese	1565
Liuro da Fabrica das Naus	Fernando Oliveira	Portugal	Portuguese	c. 1580
Livro de Traças de Carpintaria	Manoel Fernandez	Portugal	Portuguese	1616
Livro Náutico	Unknown	Portugal	Portuguese	1575-1625
Livro Primeiro de Arquitectura Naval	João Baptista Lavanha	Portugal	Portuguese	c. 1600
Memorial das Várias Cousas Importantes	Unknown	Portugal	Portuguese	1575-1625
De Nederlandsche Scheeps-bouw-konst Open Gestelt	Cornelius van IJk	The Netherlands	Dutch	1691

This includes 465 illustrations from the treatises along with approximately 3,500 images of texts. Photographs and drawings documenting The *Pepper Wreck* excavation are approximately 1,350.

Some statistics of *glosShip*, the glossary of nautical terms:

The glossary contains terms in twelve languages, in parenthesis is the number of terms in each one: English (3223), Portuguese (481), Spanish (419), Dutch (2477), Italian (32), Latin (74), French (873), Catalan (131), German (87), Danish (80), Greek (209), and Venetian (6).

ontoShipDS includes 240 classes, 70 object properties, 53 data properties, and 57 ships. In addition, there are 265 spellings and 1368 synonyms in various languages. Definitions for the top four languages are: English (1689), French (362), Portuguese (160), and Dutch (117).

APPENDIX C

ENGLISH STOP WORD LIST

The following terms are important for describing spatial properties of ship components; however, they are included in the Lucene's English stop word list. This case illustrates the necessity of modifying the original stop word list for preventing important terms in a particular domain from being left out in the indexing process, and making them unavailable for search purposes.

- at
- by
- about
- against
- between
- into
- through
- before
- after
- above
- below
- to
- from
- up
- down
- in
- out
- on
- over
- under

APPENDIX D

DTD FOR TRANSCRIPTIONS

The Document Type Definition (DTD) for transcriptions of treatises in XML is defined as follows:

```
<!DOCTYPE nadl_text [
<!ELEMENT nadl_text (folio)>
<!ELEMENT folio (graphic, l*)>
<!ELEMENT graphic (#PCDATA)>
<!ELEMENT l (text+, glossary*, add*, del*, note*)>
<!ELEMENT text (#PCDATA)>
<!ELEMENT glossary (#PCDATA)>
<!ELEMENT add (#PCDATA)>
<!ELEMENT del (#PCDATA)>
<!ELEMENT note (#PCDATA)>

<!ATTLIST nadl_text folio CDATA #REQUIRED>
<!ATTLIST graphic url CDATA #REQUIRED>
<!ATTLIST graphic mimeType CDATA #REQUIRED>
<!ATTLIST glossary id CDATA #REQUIRED>
<!ATTLIST add id CDATA #REQUIRED>
<!ATTLIST add place CDATA #REQUIRED>
<!ATTLIST del id CDATA #REQUIRED>
<!ATTLIST del type CDATA #REQUIRED>
<!ATTLIST note id CDATA #REQUIRED>
]>
```

VITA

Carlos A. Monroy Cobar received his Licentiate (Bachelor of Science) degree in computer science from Universidad Rafael Landívar in Guatemala in 1996. He joined the computer science program at Texas A&M University in September 1999. In May 2002, he received his Master of Computer Science degree. He completed his doctoral studies in computer science at Texas A&M University in August 2010. His research interests include digital libraries, digital humanities, information visualization, genetic algorithms, games for education, and the use of information technology in multidisciplinary/multicultural collaboration.

Dr. Monroy can be reached at 301 Harvey R. Bright Building, College Station, TX, 77843-3112, or by email at cmonroy@csdl.tamu.edu.