Copy No. _____

Defence Research and Development Canada
Recherche et développement pour la défense Canada

DEFENCE R&D DÉFENSE

Implicit Trust in a Data Model

Marie-Odette St-Hilaire
Michel Mayrand
OODA Technologies Inc.

Anthony Isenor
DRDC Atlantic

Prepared by:
OODA Technologies Inc.
4891 Av. Grosvenor
Montreal, QC H3W 2M2

Project Manager: Anthony W. Isenor, 902-426-3100 ext. 106
Contract Number: W7707-115137-Callup1
Contract Scientific Authority: Anthony W. Isenor, 902-426-3100 ext. 106

Defence R&D Canada – Atlantic

Contract Report
DRDC Atlantic CR 2011-107
October 2011

Canada

This page intentionally left blank.

Implicit Trust in a Data Model

Marie-Odette St-Hilaire
Michel Mayrand
OODA Technologies

Anthony W. Isenor
DRDC Atlantic

Prepared by:

OODA Technologies Inc.
4891 Av. Grosvenor, Montreal Qc, H3W 2M2

Project Manager: Anthony W. Isenor 902-426-3100 Ext. 106
Contract Number: W7707-115137-Callup1
Contract Scientific Authority: Anthony W. Isenor 902-426-3100 Ext. 106

Defence R&D Canada – Atlantic

Contract Report

DRDC Atlantic CR 2011-107

October 2011

Principal Author

Marie-Odette St-Hilaire

Approved by

Francine Desharnais
Head/Maritime Information and Combat Systems Section

Approved for release by

Calvin Hyatt
Head/Document Review Panel

Abstract

As an assessment of databases used to store Maritime Situational Awareness data, the hypothesis is posed that *a database built upon a data model, utilizing international standards, recognized and accepted data modelling concepts, best practices, etc. would be more trusted by the community utilizing the data contained within the database*. In this work, the validity of this hypothesis was investigated and assessed by decomposing trust into the sub-components: predictability, dependability, faith, reliability, robustness, familiarity, understandability, explication of intention, usefulness, competence, self-confidence, and reputation. An analysis of how these components are expressed in the context of a database system and in particular, how they impact the data model, was performed. The analysis indicates that reliability, understandability, usefulness, familiarity and reputation are the components that capture the concept of trust in a data model. These components were then applied in an analysis of the National Information Exchange Model-Maritime data model, essentially grading the model against the applicable trust components. Results vary from a poor grade on aspects of reliability, to excellent in terms of familiarity and reputation.

Résumé

L'évaluation de bases des données utilisées pour stocker des données de Connaissance de la Situation Maritime (CSM) peut se faire selon plusieurs critères. Dans ce travail, nous explorons l'angle de la confiance en posant l'hypothèse suivante : Une communauté scientifique a tendance à avoir plus confiance en une base de données conçue avec un modèle de données développé à partir de standards internationaux, des méthodes reconnues et des bonnes pratiques de modélisation de données. Cette étude explore et valide cette hypothèse en décomposant le concept de la confiance en sous-composantes : prévisibilité, sureté de fonctionnement, croyance, fiabilité, robustesse, familiarité, compréhension, explication de l'intention, utilité, compétence, confiance en soi et la réputation. Ces composantes sont définies dans un contexte systèmes de bases de données. De plus, une analyse de leur impact sur la perception d'un modèle de données est conduite. Cette analyse indique que la fiabilité, la compréhension, la familiarité et la réputation sont les composantes qui caractérisent le mieux le concept de la confiance dans un modèle de données.

This page intentionally left blank.

Executive summary

Implicit Trust in a Data Model

Marie-Odette St-Hilaire, Michel Mayrand, Anthony W. Isenor; DRDC Atlantic
CR 2011-107; Defence R&D Canada – Atlantic; October 2011.

Background: The compilation of data to support Maritime Situational Awareness can take on many forms. In a military context, data models such as Joint Consultation Command and Control Information Exchange Data Model, Maritime Information Exchange Model (MIEM), Universal Core, or the National Information Exchange Model (NIEM) claim to address similar requirements, but in some sense ”better” than the others. These conflicting claims confuse the user community and often do not progress the underlying issue of data use.

Diversity in data models is a recognized result of the data modelling activity. However, there may be many factors that influence the selection of one model over another. One factor is trust, specifically the trust the user places in the overall system. Since the data model is part of the system, it is recognized that the data model could have an impact on user trust. As a result, the Maritime Information and Knowledge Management group at DRDC Atlantic has contracted a study to investigate the relationship between the factors influencing trust, and a data model.

Results: An analysis on how to define trust in a data model was conducted under contract by OODA Technologies over a period of four months, in support of two Applied Research Projects: 11HL *Technologies for Trusted Maritime Situational Awareness* and 11HO *Situational Information for Enabling Development of Northern Awareness*. It was found that a data model can influence one’s trust in the data delivered from the system. The applicable trust components were then applied to the National Information Exchange Model (NIEM)-Maritime data model, this being the result of harmonizing the MIEM into the NIEM. The results highlight the strengths and the weaknesses of the NIEM-Maritime from a trust perspective.

Significance: The defence and security community must be prepared and able to properly assess the trust it places in a system’s delivery of data to the decision maker. This study provides an understanding of how the data model influences the user’s trust in the system.

Future Plans: Support to the Canadian Forces and their role in public security, is an evolving function for the defence community. Part of the ongoing research in this area will involve investigations of systems that will bridge traditional defence roles, such as Maritime

Situational Awareness (MSA), and security roles, such as law enforcement or more generally public safety. Such investigations will develop and strengthen the cross-departmental relationships that contribute to an integrated Canadian security environment.

Sommaire

Implicit Trust in a Data Model

Marie-Odette St-Hilaire, Michel Mayrand, Anthony W. Isenor ; DRDC Atlantic CR 2011-107 ; R & D pour la défense Canada – Atlantique ; octobre 2011.

Contexte : La collecte de données et de renseignements pour appuyer la Connaissance de la Situation Maritime (CSM) peut prendre diverses formes. Dans un contexte militaire, les modèles de données tels que le Joint Consultation Command and Control Information Exchange Data Model, Maritime Information Exchange Model (MIEM), Universal Core, ou le National Information Exchange Model (NIEM), prétendent tous répondre aux mêmes besoins, et ce « mieux » que les autres. Ces déclarations confondent la communauté d'utilisateurs, ce qui contribue peu à la solution du problème de l'utilisation de des données. Une conséquence bien connue de l'activité de modélisation de données est la diversité de ses résultats. Il existe cependant plusieurs facteurs influençant la sélection d'un modèle de données. Un de ces facteurs est la confiance, en particulier la confiance qu'un utilisateur place dans le système de base de données résultant. Le modèle de données faisant partie de ce système, il est reconnu qu'il pourrait avoir un impact sur la confiance de l'utilisateur. En conséquence, le groupe de Gestion de l'Information et du Savoir Maritime (GISM) à RDDC Atlantique a contracté une étude investiguant la relation entre les facteurs influençant la confiance et le modèle de données.

Résultats : Une analyse sur la définition du concept de confiance dans un modèle de donnée a été réalisée sous contrat par OODA Technologies sur une période de quatre mois, en appui aux deux projets : 11HL Technologies assurant la fiabilité de la connaissance de la situation maritime et 11HO Information situationnelle permettant le développement de la connaissance du Nord. Il a été constaté que le modèle de données peut influencer la confiance qu'un utilisateur porte dans les données fournies par le système. Les composantes de la confiance pertinentes pour un modèle de données ont été appliquées au modèle du National Information Exchange Model (NIEM)-Maritime, ce modèle étant le produit de l'harmonisation du MIEM avec le NIEM. Les résultats mettent en évidence les forces et faiblesses du NIEM-Maritime dans une perspective de confiance.

Importance : La communauté de la défense et de la sécurité doit être préparée et en mesure de bien évaluer la confiance qu'elle place dans un système de données utilisé pour la prise de décision. Cette étude permet de comprendre comment le modèle de données influence la confiance de l'utilisateur dans le système.

Perspectives : Le soutien aux Forces Canadiennes et à leur rôle dans la sécurité publique est une fonction de la communauté de la défense qui est en constante évolution. Une par-

tie de la recherche en cours dans ce domaine impliquera les systèmes qui feront le pont entre les rôles traditionnels de la défense, tels que la Connaissance de la Situation Maritime (CSM), et ses rôles de sécurité, tels que l'application des lois, ou plus généralement la sécurité publique. Les résultats de ces recherches vont aider à développer et renforcer les relations interdépartementales, contribuant ainsi à un environnement de sécurité canadienne intégrée.

Table of contents

List of figures

List of tables

1 Scope

Maritime information related to ship activities is a key component of Maritime Situational Awareness (MSA). In many cases, the data that support MSA are stored in databases. These databases are accessed directly by users or by software applications (automated or not) that utilize the data in some manner.

Here, we are interested in the assessment of the data models that would be used to construct the databases used in MSA-related investigations. The data content of the database obviously plays a large role in the usefulness of the database system to the MSA activity. Obviously, if the system does not contain the proper data, the user will be unable to use the system in the MSA activity. For this investigation, we are going to assume that the database contains the proper data for the user's MSA activity. Given that assumption, the database that contains the data may still take many different forms. For example, the database may consist of a small number of tables, have few internal relationships, have minimal structures that are based on international information standards, etc. Alternately, the database may be based on an extensive data modelling exercise, contain normalized tables, extensive internal relationships, metadata as defined by an international standard, etc.

As an assessment of the database structure used to store MSA data, we pose the hypothesis that

> A *database built upon a data model*, utilizing international standards, recognized and accepted data modelling concepts, best practices, etc. *would be more trusted* by the community utilizing the data contained within the database.

This document includes an investigation and assessment of this hypothesis. The document is broken into the following sections:

- Section 2 provides an overview of the literature related to trust. This section describes the components of trust.
- Section 3 provides an overview of the components of a database system. This section describes the assumed components that make up a system.
- Section 4 then describes the factors that influence trust in a database system. These factors are related to the user's perception of the system, and how the user utilizes or wants to utilize the system. This section describes the user's view onto the system, and it is from this view that user perceptions are formed.
- Section 5 describes the methodology used to evaluate the hypothesis.
- Section 6 focuses on the data model, and describes the metrics used for assessing trust in a data model.
- Section 7 describes the relative importance of the trust components and how the trust in a data model may impact the trust in the database constructed from that model.
- Section 8 then presents an example of assessing a data model in terms of the important trust components. This example uses the NIEM-Maritime data model.

– Finally, section 9 presents concluding remarks by combining the results from previous sections in the form of guidelines for developing a data model that would exhibit trustworthy characteristics.

2 Trust

Literature about trust in technologies is vast. There is considerable research on modelling trust and defining trust, its prerequisites, components, and consequences, for a wide range of contexts. Providing extensive background on trust in technologies is outside the scope of this study. For that, refer to Adams et al. [1] or Artz and Gil [2].

It is widely accepted that trust is not a binary concept; it consists of several components (or dimensions). For example, consider the concrete case of Wikipedia [1]. Among the plethora of information websites, why do people trust Wikipedia? Aside from its reputation and reliability, Wikipedia offers information that can be updated at any time (i.e., timely) within a simple construct. Moreover, there is typically metadata within each article to support the credibility of the information. Metadata takes the form of notes, references, internal and external links, timestamps, and flags, to inform the user about information quality. These are some of the characteristics that influence the levels of trust in Wikipedia. These system characteristics are part of, or related to, components of trust in technologies found in the literature.

For this study, we define trust in a computerized system as being made up of 10 components:

1. Predictability, Dependability, Faith (these exist on a continuum)
2. Reliability
3. Robustness
4. Familiarity
5. Understandability
6. Explication of intention
7. Usefulness
8. Competence
9. Self-confidence
10. Reputation

These trust components were compiled based on a literature review conducted by Isenor et al. [3].

An apparent challenge resides in the quantification and measurement of these components. For instance, to quantify Wikipedia's reputation or its understandability is not a simple task. Such quantification would require experimentation with human subjects and well-defined protocols. Since that kind of experimentation is not in the scope of this study,

1. http://www.wikipedia.org/

most components will have to be broken down into lower level components that allow measurement for the context of interest. For instance, system usefulness can be assessed using the characteristics of maintainability, integration, relevance, etc.

The remainder of this section provides the definition of each trust component mentioned in the list above. These definitions are extracted from Isenor et al. [3].

2.1 Predictability

Predictability is described as the estimation or anticipation of an outcome given a specific and understood set of inputs.

The system component must perform functions in such a way that the user would expect certain results. This expectation goes beyond the expected function provided by the system, to a prediction of how those functions operate and what they provide as output. This is obviously related to familiarity of the user to the functions. A user who is highly skilled and familiar with a particular process will have higher expectations from a system component to help them perform that function. The user will form opinions on the system component when they are able to consistently produce results based on consistent levels of input. If the system component decisions and actions are competent as compared to the level of user competence, then the user will develop positive opinions regarding the system component.

2.2 Dependability

Dependability is seen as a continuation from predictability. Predictability deals with the specific outcome from a function, while dependability deals with confidence that the function will be executed. Note that this is only confidence in execution; not necessarily confidence in the result.

2.3 Faith

Faith is seen as a continuation from dependability. It is described as the belief in a system's ability to provide a satisfactory outcome in a situation that has not previously been experienced.

2.4 Reliability

Reliability means repeated and consistent functioning. It is considered a functional property, one that indicates readiness-of-use or readiness-to-respond. The system component must be seen from the user perspective as being available and responsive to the required

activity. If the system component is repeatedly not functioning when it is required, the user will develop processes that neglect including the system.

Reliability is also linked to the expertise the user has with the system, linking user training or expertise to the user's reliance and therefore trust in the system. The Kelly et al. [4] model also relates perceived reliability of the system to the user's experience. This seems to indicate that user perceived reliability is established through direct interaction.

2.5 Robustness

Robustness is described as the systems ability to perform under a variety of circumstances (e.g., faulty hardware, busy network, low memory, slow CPU, etc.). This is important in a military domain where factors such as the available bandwidth or available data may affect system performance.

2.6 Familiarity

Familiarity is described as the adherence of the system to the procedures, terms, and cultural norms of the user. This places a direct relationship between familiarity and the user's education, experience, and training.

Familiarity has an important consequence for users trained in different methods and symbology. Users brought up through the different services often have different terminology and symbology for very similar things. For example, *tank* to a navy person would mean a container for holding fluid. To an army person, *tank* could mean a mobile gun.

2.7 Understandability

This is how the user perceives their ability to comprehend the system. It is the user's ability to form a mental model and predict future system behavior. If the user feels that he/she comprehends how the system will react and how the system can help them, then understandability increases.

Understandability is not the same as familiarity, though familiarity helps understanding (note that Kelly et al.[4] places familiarity as a sibling under understandability). Sheridan [5] proposed the following analogy to explain it: *we are all familiar with people who are not understandable or predictable, and we are not sure whether to trust them or not*.

2.8 Explication of Intention

Sheridan [6] initially defined explication of intention as the system's ability to explicitly display or say that it will act in a particular way (as contrasted to its future actions having

to be predicted from a model).

This captures the notion of explicit information being provided from the system to the user, which describes what the system is or will be doing. This has also been described as transparency. Essentially, this accounts for the system describing its actions to the user, effectively educating the user on its functions and procedures.

2.9 Usefulness

This is how helpful the system is to the user. If the system provides functionality that is either non-existent elsewhere or non-existent in a timely fashion, then the system could be considered useful to the user.

2.10 Competence

This is related to the use of the system component within the larger system. If the system is perceived as properly using the component, and if appropriate methods or procedures are employed, then the system is considered competent.

2.11 Self-Confidence

This is the user's confidence in their ability to perform the functions provided by the system. Self-confidence was found to be related to trust in a simple way for low-automation systems. In such cases, when the user's self-confidence was low, their trust in the automation was high. Similarly, users with high self-confidence tend to perform duties themselves, indicating low trust in the automation.

This indicates a rather obvious point, namely, that a user's ability to trust a system is partially defined by their knowledge of the functionality provided by the system.

2.12 Reputation

This aspect deals with the reputation of the creator, producer or maker of the system. If the user has knowledge of other's past experience with the maker, then the reputation of the maker will affect the user's initial level of trust. The importance of reputation may decrease as direct user experience increases. Thus, reputation has initial importance but the level of importance decreases with experience.

2.13 Others

The above trust components are the ones that most often appear in the literature. Other trust-related terminology appears closer to human science rather than digital processing, or included in one of the above components, or less important. The following is a non-exhaustive list of other terminology obtained from these sources [1, 7, 8, 9]:

1. Accuracy: the extent to which the system provides output free of error.
2. Power/Control: the extent to which the user is able to control the behavior of the system.
3. Adaptability: the degree to which the system can change according to a situation.
4. Experience: based on the specific user's past encounters with the system.
5. Integrity: the extent to which the system is able to recover from technical failures or user errors without loss of data.
6. Solidarity: the degree to which the user perceives how the system shares a similar purpose to himself.
7. Performance: in regards to the overall human-machine system performance.
8. Fiduciary responsibility: the degree to which the user expects that the system will meet its design-based criteria.
9. Personal attachment to the system: comprised of *liking* and *loving*. *Liking* means the user finds using the system agreeable; essentially it suits their taste. *Loving* means the user has a strong preference for the system, is partial to using it, and has some level of attachment to it.
10. Openness: User's mental accessibility, or the willingness to share ideas and information freely with others.
11. Cooperation: Importance of proximity, communication and interaction; dependability; benevolence.
12. Risk of a situation: Dependability under risky conditions; sharing of confidential information; calculative process.
13. Complexity/Simplicity: refers to the model containing the minimum possible entities and relationships. This concept has been included in understandability.
14. Integration/Implementability: refers to the ease of implementation of a model including such things as being implemented on time, on budget, and within technology constraints.

3 Database System

A database is usually part of a broader automated system. If not, the database can be considered as a part of a Database Management System (DBMS) (with the possible exception of flat file databases).

For the purpose of this study, we start by defining a database, database management system and database system.

Database: Collection of information organized in such a way that a computer program can quickly select desired pieces of data.

DBMS: Collection of programs that enable you to store, modify and extract information from a database.

Database system: Database + DBMS.

In order to determine the factors influencing the trust a user places in a data model and a database system, we need to isolate the database content (i.e., data) from the structure (i.e., data model), the access (i.e., DBMS), and the storing components (i.e., hardware) of the system. It is important to clearly differentiate the components involved in a database system in order to identify the role of the data model. To do so, we use a simple representation of an information system.

This section is organized as follows: 3.1 provides a simple representation to describe the components of any information system (including a database system); 3.2 presents a simple classification of users of the database system; and 3.3 discusses database modelling.

3.1 Components of a System

Several models exist for information systems, depending on the type and architecture chosen. For this study, we take a simplistic approach, proposed by Isenor et al. [3], that quantifies the many individual parts of an information system into four broad components. These four components were identified by considering the entire system in a networked enabled operation (see Figure 1).

First, there is the component that deals with the collection and persistence [2] of data. This is referred to as the source. Second, there are the data themselves. Third, there is the processing associated with the data. The fourth component results from the fact that the processing does not need to be conducted at the source. Thus, a delivery mechanism is responsible for the transport of the data to the processing functions.

2. In the context of this study, persistence is used as a synonym for storing.

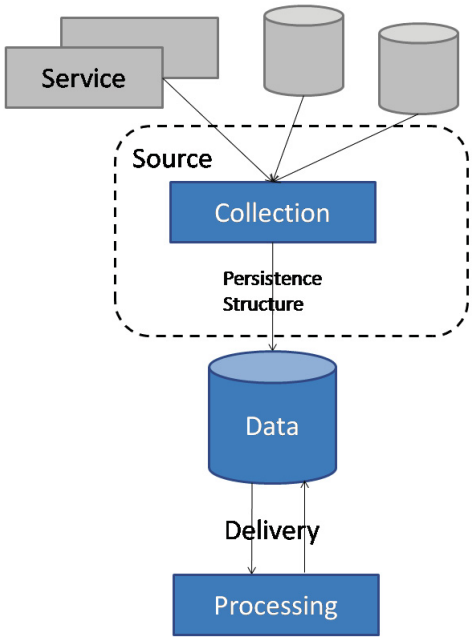

Figure 1: *Components of a system: source, data, delivery and processing.*

3.1.1 Source

For this discussion, the source is considered a self-contained entity that provides one or more resources to the broad networked community. A source is typically a collection of people and equipment that work as a contained and cooperative group. The source is likely (but not necessarily) under one management control mechanism.

A source has three main capabilities: Information collection (and potentially processing), structuring (or organization), and persisting. From this definition, a database appears as a source.

3.1.2 Data

Data represent one possible resource that can be made available by the source. The data are the numbers or characters that represent the observations, measurements, and information that is contributed by the source.

The data provided by the source must be the responsibility of that source. However, the data does not have to originate from that source. For example, a data collection activity can take place under the direction of one source; called source A. Source A can provide the data to the network. Source B can acquire the data, perform value added functions to the data, and present a new data set to the network. In this case, Source B is responsible for the value added data product while Source A remains responsible for the original data set.

3.1.3 Delivery

The delivery is the component representing the transport of the data from the source. It is the medium in which the data move.

For a database system, the delivery component can take several forms. Among these: local Application Programming Interface (API), web service/interfaces, or simply DBMS access functionalities.

3.1.4 Processing

Processing represents the analysis or manipulation of the data performed at the receiver. There will certainly be processing at the source, but here *processing* refers specifically to that processing external to the source. Processing could be performed on a single data set, or could combine (e.g., fuse) multiple data sets. As well, processing includes those functions required to discover data assets.

Processing is not part of a database system. In the context of this study, processing components are applications built upon a database system such as administration tools, data mining/analyzing tools, visualization applications, etc.

3.2 Database System Users

For the purpose of this study, we define two types of users: the scientist and the operator. See Figure 2, which is based on a figure from [10], for an illustration of the two kinds of users and their interaction with data. In this document, the term *user* refers to both scientist and operator. The term scientist or operator is used when a distinction is needed.

3.2.1 Scientist

The scientist is a *database literate* user. The scientist may be a database administrator or a developer (i.e., a term that includes investigating the signals contained within the data). In both cases, a deep understanding of the data and its structure is required to perform their tasks.

A database administrator installs and maintains databases. Their role is to manage the integrity, security, and overall performance of a database system, including tasks such as scheduling backups, database reorganization activities, optimizing storage layout, tuning performance in response to application requirements, and managing user access to the data. Depending on the company, a scientist and/or database administrator may be responsible for inserting and maintaining the metadata.

The developer uses the database system to develop knowledge, theories, or applications by utilizing the data contained within the database. The developer is part of the team who defines the application requirements, designs a solution, and implements application code to meet those requirements.

3.2.2 Operator

The operator is someone who relies on the data to make robust decisions. They usually have access to a particular view of the data and a limited part of the database. They have to interpret some aspect of the data and its structure to perform their tasks, but they do not have to be familiar with the underlying database or with data modelling.

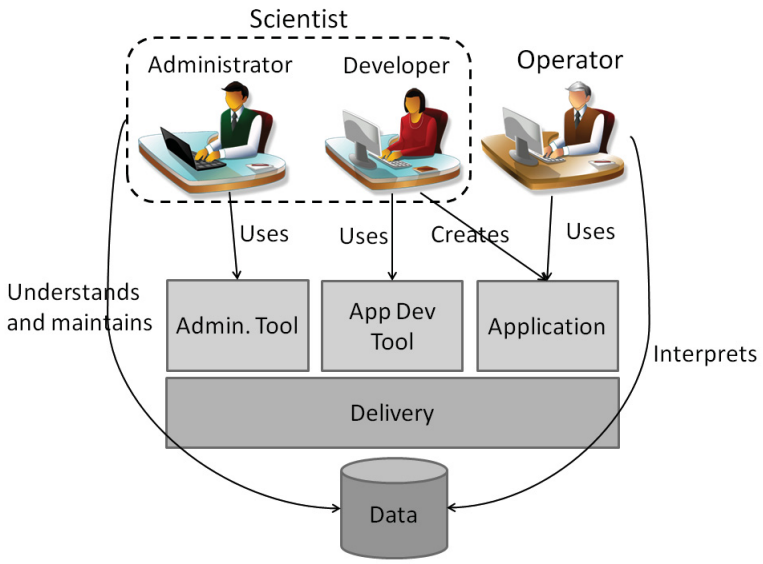

Figure 2: *Two types of database system users are defined: the scientist and the operator.*

3.3 Database Modelling

Database modelling is often called data modelling, because the resulting model represents how the individual datum are related. The resulting data model is a type of blueprint for the construction of a database. Although both database modelling and data modelling are acceptable, in this document we use the terminology *data modelling*.

The data model or schema represents the structure or inter-relationships of the data, whereas the data are the *facts* contained within the structure. The term *model* includes the *rules* that the data must obey to comply with the inter-relationships.

There are three levels of data modelling: conceptual, logical and physical. Here are summarized versions of the Simsion [11] definitions (reported in Isenor and Spears [12]):

Conceptual data modelling: a database independent view of the data

Logical data model: converts a conceptual data model into a form that uses the data definition language of a specific database implementation.

Physical data model: converts the logical data model into an implementation for a specific DBMS, where alterations can be made to address performance issues.

3.3.1 Data Model Quality

The benefits of a good data model are [13]:

Communication: A data model is the medium used by project team members to communicate with one another. The data model provides a common understanding of the data and rules related to those data.

Precision: Terms and rules in a data model are unambiguous. This means the terms in the data model can be defined and described to be specific and unique to the model, and that the rules can be applied in an automated and consistent manner to all the data held by the resulting database.

Figure 3, from West [14], illustrates how a quality data model can deliver benefit to an organization. The data model supports the data and computer systems used in the organization. This support is realized through data definitions and formats, and decision support, which drive the business. If this is done consistently across systems, then compatibility of data can be achieved. If the same data structures are used to store and access data, then different applications can share data [14], resulting in reduced costs and reduced risk to the business.

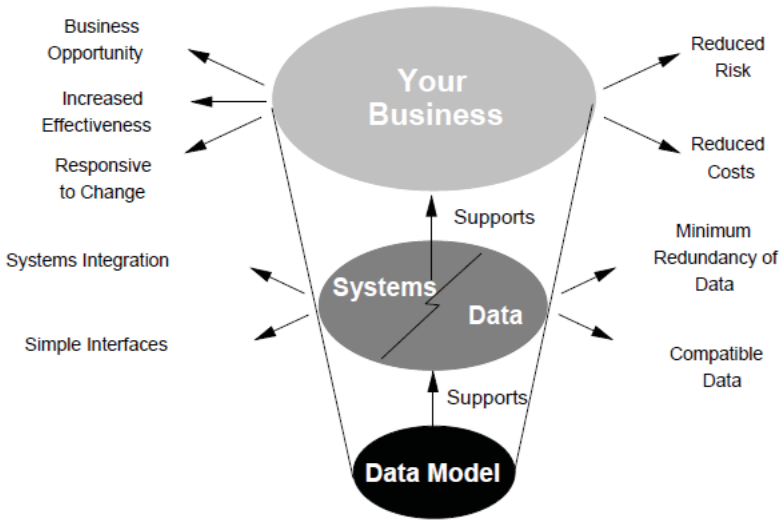

Figure 3: *How quality data models deliver benefit.*

Even data that meet all quality criteria can be useless if the data are based on a deficient data model [15]. This observation is the motivation for the research regarding the quality of a data model. Several experiments and theoretical work have been performed to define data model quality metrics (see Isenor and Spears [12] for a recent overview of the topic). However, results indicate inconsistencies in the research findings for data quality metrics and an inability to adequately measure existing quality metrics [12].

3.3.2 Standards

The direct relation between using standards and best practices in data modelling and the resulting data model quality has been documented by West [14, 16]. Figure 4, from [14], identifies standards relevant to data modelling activities. In this figure, the logical data model includes the conceptual model. Standards involved in the logical data model design are: standard context, analysis standards, and naming standards.

Standard context refers to all standards strictly related to the domain represented by the data model: industry/community standards, international standards, departmental standards, etc. Analysis standards are related to the rules governing entity and relationship definitions. These standards apply to the broader aspect of data modelling. Naming standards are concerned with the naming used to identify elements in the data model. At the level of the physical design, standard attribute formats are applied to define rules for attribute formats.

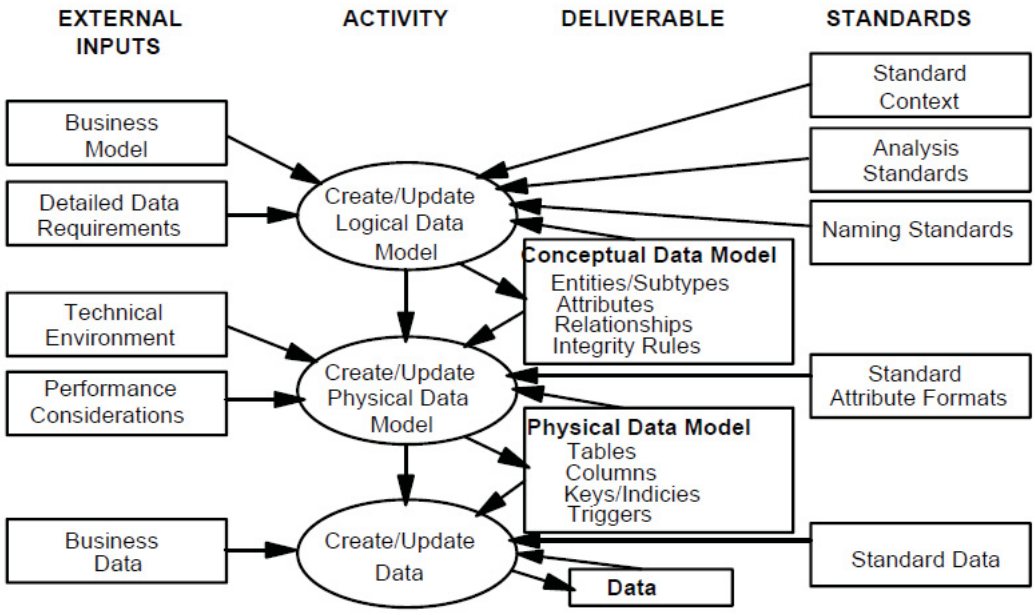

Figure 4: *Standards input into data modelling process.*

Several international data models have been developed proposing standards for information

exchange. Among these Joint Consultation Command and Control Information Exchange Data Model (JC3IEDM) and NIEM and its subset Maritime Information Exchange Model (MIEM) were developed for a military context. The last version of MIEM, which was integrated into NIEM, is analyzed in section 8 in the context of trust in a data model.

4 Main Factors Influencing Trust in a Database System

The goal of this study is to determine if a data model influences the level of trust a user places in the database system. In this context, we assert that trusting the data within the system, does not necessarily mean trusting the database that contains the data. This may be stated as:

$$\text{Trust in data} \not\Rightarrow \text{Trust in database system} \tag{1}$$

This assertion is based in part on the system components, and the fact that the data is only one part of the overall system (see Figure 1). How trust is developed by the user depends on how the user interacts with the system. For example, a user acquiring data directly from the delivery component will have different trust requirements than the user accessing data after those data have been processed. In fact, all system components implied in the data access and the user's context are influencing the user's level of trust.

Conversely, if someone claims to fully trust the database system it implies that they trust the subparts of the system; including the data within the system. This can be summarized by the following relation:

$$\text{Trust in database system} \Rightarrow \text{Trust in data} \tag{2}$$

Figure 5 illustrates the main factors influencing trust in a database system. Even if the data contained in the database are highly trustworthy, the factors (i.e., see blue oval) modify the user's perception of the data, affecting their trust in the data and in the whole system. As illustrated, the database system components, the user's context (tasks and environment) and their interaction, are like a lens modifying the user's perception of the data and thus the user's trust in the data. This illustration is inspired by the model of human trust in automation using the Lens model (see [17] for more details on the model).

The following sub-sections give an overview of the main factors influencing trust in a database system. These sub-sections use the database system and component definitions provided in section 3. Each of the factors influencing trust in the database system (i.e., see oval in Figure 5) are described.

4.1 Users, Tasks, and Environment

Trust may vary depending on who uses the database system, for what purpose, and in what kind of environment. Depending on the tasks to be executed with the data, and the user's

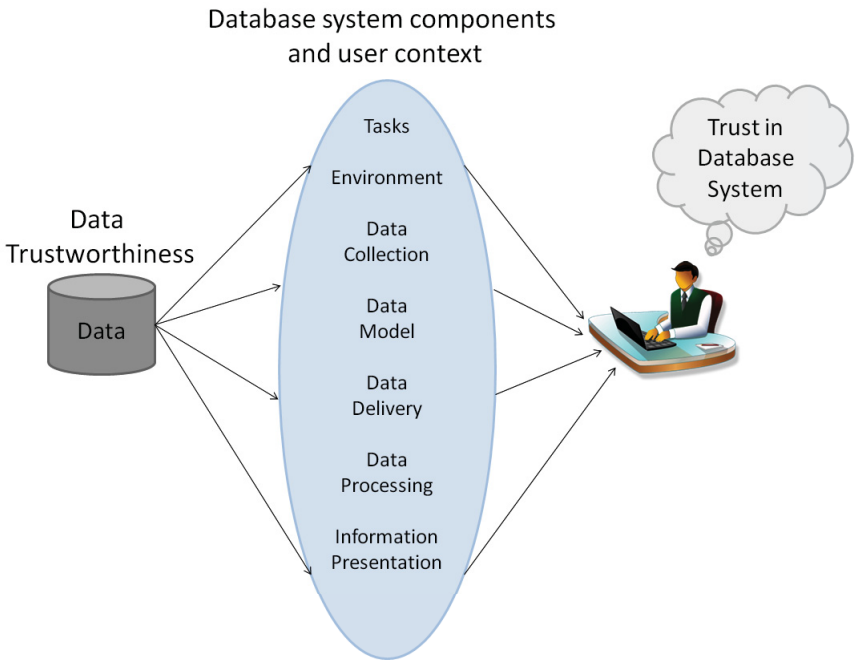

Figure 5: *Main factors influencing trust in a database system: database system components and user's context (task and environment). Even if data is trustworthy, these factors can modify the user's perception of the data and the database system.*

background (knowledge and experience), the process of building trust in a database system may differ.

As described in section 3.2, interactions with the database system will vary depending on whether the user is a scientist or an operator. Therefore, the two kinds of users will have different requirements and priorities regarding trust in the database system.

Although it is likely that all factors are important to the scientist, the data collection, data model, and data delivery are likely paramount. Scientists would also be concerned with the maintainability and integration aspect of the system, which influences trust.

For operators, the health and robustness of the database and the applications that use the database are critical; success and profitability often depends on the reliability and quality of the information produced by the applications [10]. Their trust in the database system will be highly influenced by the performance of the processing and delivery components, because the operator directly interacts with these components. In this case, trust is also directly related to the perceived information quality.

Moreover, issues of trust are more likely to become evident in high risk [3] rather than in low risk environment [1]. Organizational factors, such as restrictive data use policies, organizational history, and training, are other examples of contextual factors influencing trust in a database system.

4.2 Data Collection

Data collection is the source capability that acquires the information to be persisted in the database (see Figure 1). Reliability and reputation of the data collection process are major factors in building trust in a database system. Since data collection often involves other data sources and processes, reputation assessment is linked to the reputation of these other entities. This reputation tracking is referred to in this document as provenance and is discussed in more detail in section 6.2.1.3.

4.3 Data

Data quality and perceived data quality is a widely studied topic, especially in decision support system development. Perry et al. [18] provides a good overview of information quality in the context of situational awareness. Data quality is usually defined has having multiple dimensions such as completeness, accuracy, uncertainty, reliability, etc.

Perceived data quality plays a central role in the trust that a user places in a database system. In fact, if a database contains data that are not trustworthy, then it is unlikely the database itself would be trusted (i.e., negation of logical relation 2: Data not trustworthy $\Rightarrow$ No Trust in database).

4.4 Data Delivery

Trust in delivery is a subject of increasing interest since the booming of web technologies. Those accessing data from a computerized system should now be asking themselves: Are these data safe for my system? Do they contain malicious software? Who else has access to these data? Have the data been tampered with?

Most of the literature about trust in an information system is about security in data delivery. It covers a broad range of topics such as database security; security in complex, evolving distributed systems; web of trust; policies; computational trust; etc.

Data delivery, described as part of a database system in section 3.1.3, affects trust mainly because of security. Robustness and reliability of the delivery mechanism are also major factors influencing the level of trust, especially in an operational context.

3. Risk is related to the consequences of errors.

4.5 Data Processing

A lot of work has been focused on determining factors influencing a user's trust in automated systems (see [1] and [4] for literature reviews on the topic). The majority of this work addresses the processing component, because this is the component with the highest automation level in the system. Since processing is usually the component the operator directly interacts with, it is natural that it significantly affects the trust in the overall database system and thus in the data.

4.6 Information Presentation

Information presentation is the way information is shown to the user: application display, web site, documentation, etc. It is related to anything that is used to visualize the information or advertise the database system and content.

There is an important body of literature about trust in web applications and presentation is always identified as a main factor of trust [4]. See Golbeck's [19] survey about trust on the web for an overview on the topic.

The literature indicates that presentation influences the perceived credibility in web content. For example, some researchers argue that user trust is primarily driven by an attractive and professional design, the presence or absence of visual anchors or prominent features such as a photograph or kitemark, and the information structure on the website. Fogg et al. [20] and Grabner-Kräuter and Kaluscha [21] provide details about the impact of presentation on trust in a web application context.

4.7 Data Model

The quality of a data model (discussed in section 3.3.1) and the impact of the data model on trust in a database system are in someway related. It is obvious that if a data model is evaluated as of poor quality, it will have a negative impact on trust. Since there is no consensus on quality metrics, we are considering these two concepts separately. The current investigation is about trust only. We are referring to the data model quality literature when it is relevant to assess trust.

It was demonstrated in [20] that the information structure in a website has a significant impact on how much a user will trust the website content. To some extent this indicates that the trust in information contained in a website is influenced by the way the information is structured. Since a website is an information container, can we establish a parallel with trust in database and data model?

4. Other factors are mainly security and reputation.

The remaining of the document is dedicated to this investigation.

5 Hypothesis Validation Approach

Trusting a database system means trusting its information collection, structuring and storing capabilities, its data and the delivery mechanism, and also the way these parts interact. The processing component and the presentation layer can also be added to the list if those are part of the system. However, we are omitting them for this study.

Therefore, trusting a database system implies trusting the data structure or more generally the data model. This makes sense from a logical point of view, but does it makes sense in reality? Also, to what extent? Is it negligible or central in the trust a user places in a database system?

As an assessment of the types of databases used to store MSA data, we pose the hypothesis that:

> A *database built upon a data model*, utilizing international standards, recognized and accepted data modelling concepts, best practices, etc. *would be more trusted* by the community utilizing the data contained within the database.

The remainder of this document is dedicated to the validation of this hypothesis. *Validation* here is an exaggeration. To validate a hypothesis means testing it with new empirical material. In this particular context, it means performing experiments with human subjects. This would involve: designing questionnaires, selecting a significant number of data models and testers, conducting the experiment and analyze the results. This approach is not possible in the context of this study. Here, the use of the word *validation* is closer to *analysis*. So we are analyzing, based on previous research, the impact of the data model on the trust a user places in a database system.

In order to validate the hypothesis, we make the following assumptions to simplify the problem:

1. there is no process component in the system,
2. the information presentation layer is not considered,
3. the data delivery (hardware/Operating System (OS)/DBMS) is totally safe and trustworthy,
4. the data is of high quality (complete, accurate, etc.),
5. the data (raw and processed) collection is trustworthy.

Here is a scenario to illustrate the problem:

Suppose a database was developed to persist the information collected by a highly trusted data collection capability. Can someone develop trust in this database? Does the data model influence the overall trust in the database? Finally: Can someone trust a data model?

Based on the assumptions above, hypothesis validation is performed with the following steps:

1. Use the trust components in section 2 and determine which ones can be used for assessing trust in the data model (sections 6.1 to 6.10).
2. Determine the relative importance of the trust components defined in step 1 (section 7.1).
3. Determine if trust in a data model influences the level of trust in the data to be contained in the database (section 7.2).
4. Validate the hypothesis (section 7.3).
5. As a test, apply the results to the NIEM-Maritime data model (section 8).

5.1 Methodology

As mentioned above, no experimentation with subjects will be performed to validate the hypothesis. Instead, we analyze the literature for evidence for or against the hypothesis.

This validation process will be supported by documented research in:
– trust in automated systems,
– trust in information technologies (including semantic web),
– data quality and data model quality.

The topic of trust in database systems lies between the topics of trust in automated systems and trust in information technologies. A database system is an automated system but with a low level of automation [5]. Findings in that domain cannot all be directly applied to databases. Indeed, special attention must be given to the role of the level of automation in trust assessment.

The study of trust in a database system also falls in the information technology area. Networked information systems usually expose databases or knowledge bases. Findings in the area of trustworthiness of networked information systems may therefore be applicable to our problem.

As mentioned in 4.3, it is a requirement to trust the data contained in a database in order to trust the database itself. That is why data quality and data model quality (closely related to data quality) are part of these research areas.

5. Level 1 on the automation scale of Moray and Inagaki [22]: The human does all the planning, scheduling, optimizing, etc. and turns task over to computer for merely deterministic execution.

6 Trust in a Data Model

It is difficult to distinguish trust in a data model from trust in the other system components, as described in section 3. As well, the user's trust is influenced by not only the system components, but also additional factors as presented in section 4. In an attempt to clarify this distinction, assumptions given in section 5 are now used to scope the trust components exclusively to a data model. In essence, the following section assesses the trust components presented in section 2, against the concept of a data model.

6.1 Predictability/Dependability/Faith and Data Model

Predictability is described as the estimation or anticipation of an outcome given a specific and understood set of inputs. It describes a reaction to an action. Since data and its structure do not provide feedback, predictability is not a natural fit for a data model.

However, anyone using a database has some anticipation about its content and structure. There are always expectations about relationships between entities, data formats, or naming standards, for instance. The user has the expectation that the data model will fit their own mental representation of data. This representation is built through experience, education and requirements for the task at hand. For example, dealing with a table of numbers representing sea surface temperature, it would be reasonable to predict that the values will be in the range of approximately -2 to 40 Celsius.

This extension of predictability to the data model is linked to understandability and familiarity. These components are also related to the user's background and experience. For that reason and the fact that predictability is related to system feedback, we will not use predictability to asses a data model.

Also, dependability and faith, being a continuation of predictability, will not be used to assess a data model.

Summary: predictability, dependability and faith, will not be used to assess a data model.

6.2 Data Model Reliability

Repeated acquisition of high quality data will instill in the user a feeling of reliability. Since data acquisition by the user is enabled by a database system, the data model plays a role in the perceived data quality and therefore a role in the user's perception of reliability.

There is a large body of literature about data and information quality. In our context, data quality from a data consumer point of view is particularly pertinent. The work of Wang

and Strong [23] on this topic is widely cited. Wang and Strong [23] define data quality as consisting of several attributes. Among these attributes, some can be influenced by the data model and among this subset of attributes, some can also affect trust.

The goal is not to assess data quality, but to isolate the role of the data model in the quality perception. To do this, we examine data model reliability by starting at the database system level. We assume that from a user perspective, repeated acquisition of high quality data from the database system will instill in the user a feeling of reliability. Thus, the data model may be considered by the user to be reliable, if the database constructed from the data model is capable of preserving and representing data quality. Thus, data quality attributes act as a type of proxy for data model reliability. From Wang and Strong [23], the relevant data quality attributes are completeness, integrity, credibility, consistency, and timeliness.

We define data model reliability as the data model's capacity to preserve and represent data quality. If we assume that data persisted in a database is of high quality, a reliable data model:
– allows the *preservation* of the initial data quality after multiple transactions and
– structures data such that its quality can be *perceived* by the database user.

6.2.1 Relation Between Data Quality Attributes and Reliable Data Model

In the following, we assume the data quality attributes may be used as a proxy for data model reliability, and thus related to the user's trust in the data model.

6.2.1.1 Completeness

Completeness is a context-dependent data quality attribute that refers to the extent to which data are of sufficient breadth, depth and scope for the task at hand [23].

A data model can influence one's perception of lack of coverage or precision of the data. In addition, a data model can lead to a loss of data completeness or make the data appear incomplete.

The following cases are examples where a data model (relational model) brings a perception of lack of coverage or precision to the data.

1. If there are too many attributes (columns) for a given relation (table).
2. If the data model granularity differs from the granularity of the source collection output.
3. If there are not enough relations to map the source output.

The first case leads to the perception of a lack of data. If there are too many attributes for a given relation, the data may appear sparse (e.g., lots of null values). For instance, consider

a database system that persists the output of a data collection process. Suppose that a given data item is reported only very occasionally by a source and was identified as not critical for the system users. Including this data item in a many-attribute table may result in the appearance of incompleteness. More refinement to the data model would likely alleviate this issue (i.e., null tracking in the data model is a common technique).

The second case can lead to the perception of a lack of precision. Here granularity of data refers to the fitness with which data fields are sub-divided. The situation where multiple data items from the collection output are merged into a single attribute in the database is a common illustration of that case.

Finally, the third case can lead to the perception of a lack of coverage. Indeed, having all data in a very restricted number of tables (big tables with many columns) may give the impression that the model is not rich enough to include all data from the source output.

6.2.1.2 Integrity

There are many definitions of data integrity, some including completeness. For the sake of this study, we use the BusinessDictionary.com [24] definition:

> Data Integrity: The accuracy and consistency of stored data, indicated by an absence of any alteration in data between two updates of a data record. Data integrity is imposed within a database at its design stage using standard rules and procedures, and is maintained using error checking and validation routines.

Data integrity is mainly achieved by preventing accidental, deliberate, or unauthorized insertion, modification, or destruction of data. Some of the techniques used are: a locking mechanism, safe transactions, and backups. Other data integrity functionalities include the implementation of integrity rules, triggers and stored procedures.

Just by examining the data model used to build a database, an experienced user can determine if there is a risk of data integrity deterioration. Indeed, a properly normalized relational database allows for:
– the efficient use of storage space
– the elimination of redundant data
– the minimization of inconsistent data
– the reduction in database maintenance overhead

Normalization is the process of efficiently organizing the data in a database. It encompasses a set of best practices designed to eliminate the duplication of data, which in turn prevents data manipulation anomalies and loss of data integrity.

Without going further into database normalization (normalization technicalities are out of the current scope), we can state that an experienced user can make a judgment about data

integrity by examining the data model and data quality deterioration. This is true, however, only for an experienced database user (i.e., a scientist). A typical operator will observe integrity deterioration but will not necessarily associate it with the data model.

6.2.1.3 Credibility

Data credibility or believability is the extent to which data are accepted or regarded as true, real, and credible [23]. A data model can structure data such that its credibility can be perceived by the database user. The common way to do so is to introduce metadata. Metadata can be used to describe:
– data quality, and
– data provenance.

Quality Metrics: Metadata can be used to qualify the information contained in the database. Quality metrics are based on the context and the information available about each piece of data. For instance, it can be related to the collection process reputation (including the reputation of all sources involved). These metrics have to be defined in collaboration with users, because the metadata is intended to inform the users about data credibility.

Since reputation is not always known and is most of the time subjective (to context and user's background), a crucial factor in credibility is provenance.

Provenance: The provenance of a piece of data is the process that led to that piece of data [25]. Here *process* includes the identity of all sources that collected that piece of information and all manipulations performed on the information (e.g., time of collection, name of expert, measurement protocol, apparatus description, unit change, rounding, combinations with other data, etc.). For this study, we consider the concepts of provenance and lineage as being the same.

Understanding the provenance of data is important for understanding the quality of data. For many scientific domains, including MSA, new databases are often created to support the data analysis needs of domain-specific scientists. Data that is collected from other sources is often cleansed and reformatted before it is compiled into the new database. Very often, the newly created database will also contain new analysis or results that are derived by scientists. By associating old and new data together in the new database, an integrated perspective is provided to users [26].

Provenance is studied for different purposes and contexts: web and linked data, Service Oriented Architecture (SOA) and web architectures, data

delivery and security, databases, etc. Provenance was identified as a critical factor of user's trust in semantic web data. The World Wide Web Consortium (W3C) created the Provenance Incubator Group in 2005 to provide a state of the art understanding and to develop a road map in the area of provenance for semantic web technologies, development, and possible standardization (see [27] for the group reports and additional references).

We consider provenance and trust to be linked, because provenance provides the information that helps users track the reputation of the sources involved in the process. For provenance in large distributed architectures, the work of Moreau et al. [25] is a good starting point and see Ceolin et al. [28] for an example of such system provenance in a MSA context. In our context of trust in a database system, the purpose of provenance is to ultimately convince the user of the quality of the data contained in a database.

The provenance of a piece of data is to be represented in a database system by some suitable documentation of the process that led to the data in the form of contextual metadata [25]. Therefore a database can be structured such that data provenance is also represented and persisted. Metadata included in the database is the common solution at the data model level. An example is the ISO 19115 process that provides the structure for gathering the metadata of geospatial data items.

Structuring data such that provenance information is omitted or not well described can affect the data quality perception (more precisely its credibility) and thus one's trust in it. At the simplest level, a data model must at least indicate if data is raw or processed.

6.2.1.4 Representational Consistency

Representational consistency is the extent to which data are presented in the same format, are compatible with previous data [23], including data types and units. The use of standards is a common way to assess this issue. Representational consistency is a quality attribute that is directly linked to the data model.

If data transformations are performed, they should be done, when it is possible, according to community standards. For instance, it is expected that gross tonnage is computed (or reported) using the standard formula provided in Regulation 3, Annex 1, of The International Convention on Tonnage Measurement of Ships, 1969 [6].

This aspect also applies to unit and type transformation (e.g., string to date object) and specification (e.g., metres or feet).

6. http://www.admiraltylawguide.com/conven/tonnage1969.html

Representational consistency is of special concern for operators. Since they base their decisions on database content, and sometimes in a stressed environment, data must be presented in the format familiar to the operator.

6.2.1.5 Timeliness

Timeliness is the extent to which the age of the data is appropriate for the task at hand [23]. A data model may or may not include time information. Depending on the tasks and data (e.g. static or dynamic data), time may be used as an indicator of quality for the data in the context of the task and the present time. Structuring the data such that age (i.e., elapsed time) is omitted or not clearly evident, would impact the user's perception of data quality. Data may not appear as sufficiently timely for the task at hand.

6.2.2 Reliability Evaluation

Here we considered data quality attributes as a type of proxy for data model reliability, and thus the evaluation was based on the data model's ability to preserve and represent the data quality attributes. These attribute are considered useful in assessing reliability of a data model.

Summary: reliability will be used to assess a data model using the proxy of data quality attributes.

6.3 Robustness and Data Model

Robustness is described as the systems ability to perform under a variety of circumstances. This is important in a military domain where factors such as the available bandwidth or available data may affect system performance.

Robustness is associated with the dynamic components of a system such as data collection and delivery [3]. It qualifies performance, which cannot be related to static components such as the data or the structure.

Summary: robustness will not be used to assess a data model.

6.4 Familiarity of a Data Model

Familiarity refers to the pre-existing knowledge that the user possesses before actually encountering the data model. It is something that is perceived instantaneously, almost at first sight. Indeed, familiarity is based on the user's past experiences and education.

Therefore, familiarity will be one of the first components to instill in the user a feeling of trust.

The data model represents the complete information requirement for the area being modeled, from the user's perspective. Since the user is always part of a community/industry, using community naming conventions or at least a common community vocabulary to model data is a way to represent the user's perspective.

Structuring data using a common community taxonomy is also a way to link the data model to the user's pre-existing knowledge. For instance, in the *tank* example given in section 2.6, the army person might be familiar with the concept of a tank object classified as an armored fighting vehicle that is a piece of equipment that is a piece of material (e.g., the JC3IEDM structure).

6.4.1 Familiarity Evaluation

Familiarity is described as the adherence of the data model to the procedures, terms and cultural norms of the user. Familiarity is evaluated by determining to what extent the data model uses vocabulary and taxonomy common to the user's community. Familiarity is fulfilled by using industry/community standards.

Summary: familiarity will be used to assess a data model.

6.5 Data Model Understandability

For a data model, understandability is different from familiarity. A data model could be developed using a vocabulary very familiar to the user but may not be understandable in terms of its structure.

Research has been conducted on defining and testing understandability metrics for a database. Metrics will differ depending on the database model selected to conduct the study: dimensional, object-relational, etc. However, in all cases, the valid metrics quantify the model complexity. Moreover, the documentation used to describe the data model plays a significant role in understanding.

6.5.1 Factors Affecting Data Model Understandability

The main factors affecting understandability of a data model are:
– structure, and
– documentation.

6.5.1.1 Structure

McGee [29] identified three properties of data models that enhance their ability to be learned and understood:

Simplicity: Refers to the number of structure types (e.g., tuples and relations) and the number of rules that govern the assembly of those structure types.

Elegance: Describes the ability to create the model using the smallest number of structure types.

Picturability: Degree to which the model lends itself to a visual representation.

These properties are directly related to complexity. Most of the literature on data model understandability postulates or demonstrates the link between complexity and understandability. The more complex a data structure is, the more user effort will be required to create a mental model of it, which may impact their ability to understand it. Serrano et al. [30] illustrated this relation as in Figure 6. Several metrics were developed for different kinds of data models (e.g., normalized, dimensional, and object-relational) and levels (conceptual, logical and physical). For a general overview of metrics to quantify the complexity aspects of a conceptual data model, see Genero and Piattini [31].

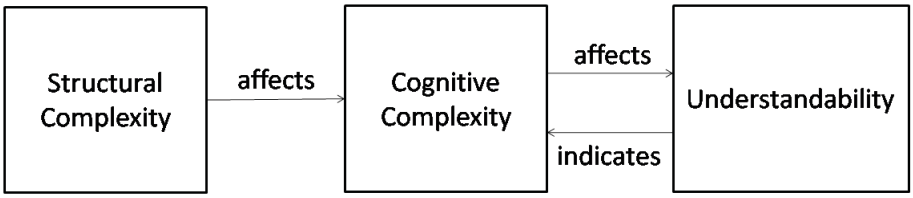

Figure 6: *Relationship between structural properties and understandability.*

From a system theory point of view, a system is called *complex* if it is composed of many different types of elements, with different types of dynamically changing relationships between them [32]. Therefore, the complexity of a data model could be highly influenced by the different elements that compose it, such as entities, attributes, relationships, generalization, etc. It is, however, impossible to get one value that captures all the complexity factors of a data model [33]. However, a rough estimate of the number of entities plus the number of relationships (for an Entity-Relationship (ER) representation) was found to be a simple and useful data model complexity metric [34]. This can be used for choosing between alternative model choices.

6.5.1.2 Documentation

As for any kind of software, documentation plays a crucial role in understanding. Documentation is not part of the data model, but it can be developed to support the data model.

In fact, documentation is a recognized part of software engineering. Documentation can take several forms: entity-relationships models, semantic models, graphical model, or plain text.

Proper documentation presents information in a manner that is easy for the reader to absorb, understand, and act upon. Its goal is to improve performance by educating users and decision makers on system capabilities or process step details. The breath and depth of the documentation produced thus depends on the audience. For someone who just uses data, like an operator, the documentation does not need to be as detailed as would the documentation for a scientist. The latter needs a deeper understanding of the data model to maintain or integrate the data held within the database that is created from the model. For instance, a conceptual model diagram may be sufficient for the operator and management team, while the logical and physical models may be required for the scientist.

Complete documentation for a relational database usually includes:

Requirements: User characteristics, objectives, problem statements, performance and security requirements.

Conceptual model: Description of entities and relationships.

Logical model: Entities, attributes, relationships, and associated descriptive comments. The comments should incorporate user community (business) names for the content of the model.

Physical model: Tables, columns, keys, data types, validation rules, database triggers, stored procedures, domains, and access constraints. It uses names limited by the DBMS and any company-defined standards.

Documentation that deals with the representation of the data model (i.e., the way the data model is communicated to the user) also plays a role in understandability. In the case of a relational database, Juhn and Naumann [35] found that graphical representations (a semantic model such as the ER model) are associated with higher levels of comprehension than tabular representations. In addition, Nordbotten and Crosby [36] concluded that graphical representations with simpler graphic styles (as opposed to sophisticated) are easier to understand. However, as mentioned by Moody [37], several researchers have noted the ER model's inability to cope with complexity. For large data models [7], complexity quickly becomes overwhelming. As a result, data models become very difficult for people, particularly non-technical users, to understand. See Moody [39] for an evaluation of alternative models for large databases. Finally, Delucia et al. [40] compared Unified Modelling Language (UML) class diagrams and ER and found that UML class diagrams significantly improve the comprehension level achieved by subjects.

7. Surveys of practice show that application data models consist of an average of 95 entities, while enterprise data models consist of an average of 536 entities [38].

Regardless of the type of documentation selected, good documentation can compensate, or at least lower, the impact of complexity on understanding. The more complex the model is, the more detailed the documentation should be.

6.5.2 Understandability Evaluation

Understandability is how the user perceives their comprehension of the data model. If the user feels they comprehend the rules that the data obey, and how the data described by the data model can help them, then understandability is reached.

User's understandability is experimentally measured with:

Comprehension: the ability to answer questions about the data model (speed, correctness).

Verification: the ability to identify discrepancies between the data model and a set of user requirements in textual form.

The following guidelines are suggested when dealing with the user's comprehension of a data model.

1. Documentation including the elements mentioned in the list above (see 6.5.1.2).
2. Data model capturing all the complexity of the problem with a minimum number of entities plus relationships.

Metrics proposed in the mentioned literature (see discussion in 6.5.1.1) can also be used to quantify complexity and thus estimate the impact on understandability.

Summary: understandability will be used to assess a data model.

6.6 Explication of Intention for a Data Model

Kelly et al. [4] places explication of intention as a sibling under understanding, while others consider them independent. Sheridan [5] made the following distinction: with explication of intention, intentions of future actions are specified outright by built-in computer-based decision, control, and automation systems; with understanding, future actions must be inferred from a deeper understanding of how the system works.

Explication of intention seems to imply that automation is specifically charged with a responsibility to provide information about its inner workings. Thinking about automation as having *intention* seems to have less meaning when thinking about simpler forms of automation [1].

Consequently, explication of intention is something that the system provides, while understandability can also refer to the state of the system.

Summary: Explication of intention will not be used to assess a data model.

6.7 Usefulness of a Data Model

This is how helpful the data model is to the user. Trust will not likely be assessed by the user, if the data model is considered useless by the user. Assuming the data contained in the database created from the model is useful to the user, its structure (i.e., the data model) can be considered useful or not.

Usefulness can be broken down into the following lower level attributes:
– relevance
– reusability
– flexibility
– cost of knowledge acquisition

6.7.1 Relevance

Data model relevance is the extent to which it is applicable and helpful for the task at hand (adaptation from data relevance in [23]).

The data persisted in a database provides a solution to a given problem. The data model represents the solution domain corresponding to the problem domain. The danger is the development of a solution domain (i.e., the data model) that does not correspond to the user's problem domain. This may occur, if for instance, the data model corresponds to the developer's perception of the user's problem domain. Indeed, designers attempt to conceptualize the problem domain into a suitable physical model subjective to many performance constraints [41]. The resulting data model can be useless to the user, even if the initial collected data was identified as helpful for the task at hand.

It is difficult to design a relevant database (with high semantic value) without significant domain knowledge and experience [41]. The key is to identify the users and involve them, if possible, in the modelling process. If it is not possible, one must find other ways of engaging the user community.

6.7.2 Reusability

Reusability is the extent to which the model allows data to be used for purposes beyond those for which the data was initially collected. Reusability includes the ease of model integration into a broader system. If the Database (DB) is built upon a common data model, interfacing will be easier [8]. This aspect is usually not a concern for operators.

8. Interfacing can account for between 25-70% of the cost of current systems [14].

Reusability can be estimated using Moody's metrics for integration [34]:
– Number of data conflicts with corporate data model
– Number of data conflicts with existing systems
– Number of data items duplicated in existing systems or projects
– Rating of ability to meet corporate needs

If the user is a scientist who needs to integrate the database system into another system or reuse it to develop applications, a data model based on community/industry standards (or at least using the same terminology) will be seen as reusable. Moreover, it is more difficult to identify potential of reuse for a data model that is not designed using data modelling standards.

6.7.3 Flexibility

Flexibility is defined as the ease with which the data model can cope with business and/or regulatory change [34]. A data model is useful if the user can extend it easily with new items and relations, i.e. if the data model is capable of evolving.

Flexibility can be estimated using Moody's metrics [34]:
– Number of data model elements that are subject to change
– Probability adjusted cost of change
– Strategic impact of change

Flexibility is realized when generic data model items are truly generic. For example, a *transportation* type object in a data model should be capable of defining ships, planes, cars, etc. If the object is not generic, then the more obscure modes of transportation that were not thought of by the designer, will not fit with the existing model. In the case of the *transportation* type, one could ask if the model could evolve to include such objects as a segway.

6.7.4 Cost of Knowledge Acquisition

Russell et al. [42] introduced cost of knowledge with the following observation:

> In a world of abundant information, but scarce time, the fundamental information access task is not finding information, but the optimal use of a person's scarce time in gaining information.

The same authors investigated the cost of knowledge in cases where the user was seeking information in what they call a *direct walk* of an information structure. They define a direct walk to be a task in which a user navigates from a starting point to a goal point in an information structure by a series of mouse points or other direct-manipulation methods [43]. The longer the walk, the higher the cost of acquisition.

This concept of direct walk can be applied to data stored in a database. If it requires multiple steps (i.e., combination of queries) to access the information that is important for the user, then the structure induces a high cost of acquisition for the given required piece of information. Conversely, if that same piece of information requires small efforts and a limited time to get, then the cost is low.

A data model structured such that the information important to the user is not straightforward to acquire, may seem less useful to that user.

6.7.5 Usefulness Evaluation

A data model is considered useful if

1. it maps to the solution of the user's problem;
2. it is easily reusable in other contexts;
3. the data model can evolve;
4. the data is easily reachable.

Items 1 and 4 are more important for operators. Items 2 and 4 are of interest for scientists. Due to associated costs, items 2, 3 and 4 are of special concern to business management.

Following data modelling best practices and standards is a safe way to produce a useful data model. Figure 7, from [14], covers usefulness components described in this section with their consequences and interactions. As illustrated, insufficient data modelling standards impact data model usefulness.

Summary: usefulness will be used to assess a data model.

6.8 Competence and Data Model

Competence qualifies the interaction with other components. This is related to the dynamic aspects of a system. Since a data model is an inert aspect of the system [3], competence cannot be used to assess the model.

Summary: competence will not be used to assess a data model.

6.9 Self-Confidence and a Data Model

A user has a level of self-confidence in their ability to interact with a system and to efficiently perform the functions provided by a system. Since the data model is inert and

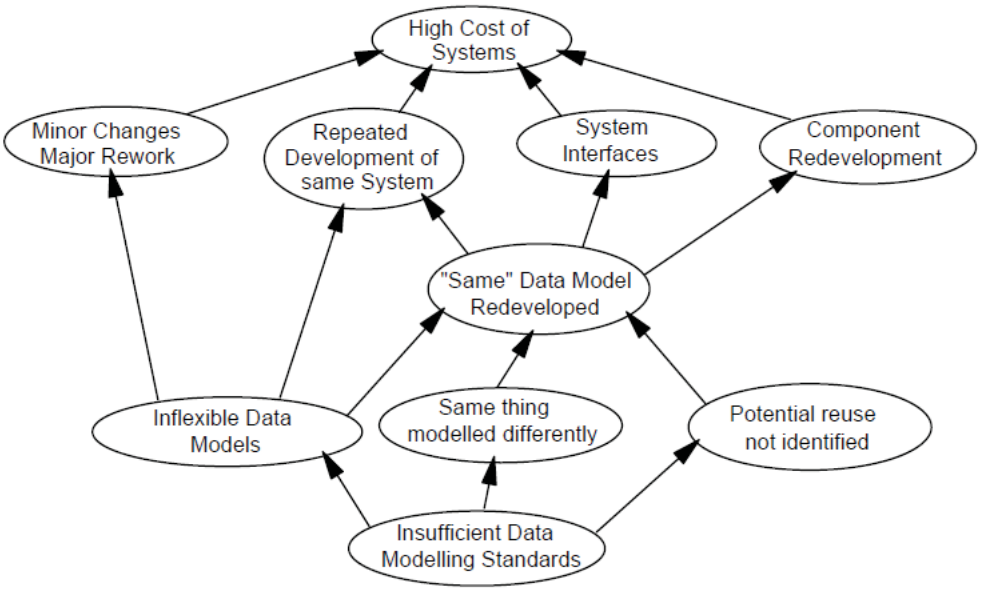

Figure 7: *Relation between data model standards and usefulness components.*

does not offer functionalities to interact with, the process of building self-confidence differs from the case of a dynamic automated system. The ability of a user to interact with a data model (i.e., get the information required) depends on their understanding of the data model. Therefore, this extension of self-confidence to a data model is strongly linked to the understanding one has in the model. For that reason and the fact that self-confidence is related to system functionalities, we will not use self-confidence to asses a data model.

Summary: self-confidence will not be used to assess a data model.

6.10 Reputation of a Data Model

The reputation of a data model is really the reputation of the creator, development team, or maker of the data model. Reputation of the maker can influence the user's initial level of trust. However, we expect the impact is relatively small. This expectation is due to the human focus on the product (i.e., the data model itself), with the reputation of the maker playing a smaller role as compared to the product.

If the user has knowledge of other's past experiences with the maker, then the reputation of the maker will influence the user's initial level of trust. The importance of reputation may decrease as direct user experience increases, or alternately may act to solidify the users preconceptions. Thus, we expect reputation to have an initial importance, but for this level to decrease with experience gained by the user. Reputation of the data model designer is more like an initial indicator, rather than a long term influence.

6.10.1 Reputation Evaluation

Data model reputation is evaluated by considering the presence of opinions about the maker and the model in the user's community. Reputation is based on the opinions of others and is much less personal as compared to direct user experiences. In cases where the data model cannot be evaluated or is unknown, it is likely that reputation has no effect on a user's trust of the data model. However, the expression of other's experiences is gaining importance. A recent study [44] indicates how social media may be changing the importance of personal past experiences compared to the opinions expressed by others. In particular, the study indicates that peer-review within the social media framework, improves the quality of a product (specifically, improves data quality). This finding indicates how social media is influencing our opinions.

The use of community standards for data modelling will contribute to the data model's reputation. Depending on the standard used, it can affect positively or negatively the trust a user places in the data model. Each data model standard claims to address similar requirements, but in some sense *better* than the others. Each development team considers their particular model to be better suited to the requirements [12]. These conflicting claims often contribute to development team and data model reputation and can influence the user's trust.

Summary: reputation will be used to assess a data model.

7 Relationship Between Trust in a Database and its Model

In section 4, we described the main factors influencing trust in a database system: users, data, delivery, etc. In section 6, we isolated the data model and used the trust components to assess a data model. By doing so, it was found that the following trust components could be used to assess a data model (not in order of importance):

Reliability: The data model's capacity to preserve and represent data quality.

Familiarity: The adherence of the data model to the procedures, terms, and cultural norms of the user.

Understandability: The extent to which the user perceives (in a positive way) their comprehension of the data model.

Usefulness: The extent to which the data model is helpful to the user.

Reputation: The reputation of the data model creator or development team.

We consider these five components as capturing the concept of trust in a data model.

The remainder of this section is organized as follows: section 7.1 describes the relative importance of these trust components; section 7.2 discusses the influence of the data model on the level of trust in a database and its data; and section 7.3 concludes by validating the hypothesis.

7.1 Relative Importance of Trust Components to Assess a Data Model

The only way to truly determine the relative importance of trust components is by experimentation with subjects. However, an acceptable alternative is to tentatively order components based on previous studies (although no documentation explicitly addressing the topic of trust applied to a data model was found). This is the approach presented here.

Several conclusions were made about trust components in the domain of trust in automation. The most recurrent results claim that competence and reliability play the major roles for building trust in automation. The widely cited work of Muir et al.[45, 46], found evidence that system competence best captured the concept of trust in a system. However, as mentioned in section 6.8, competence cannot be used to assess a data model. Competence qualifies the system's interaction with components and its aptitude to perform its tasks, which are related to the dynamic aspects of a system. This underlines the fact that research

results related to trust in automation cannot always be applied to data models. These results usually describe the trust building process in systems with a high level of automation, which is not the case of data models nor database systems.

As mentioned by Adams et al. [1], research indicates that reliability is a strong predictor of trust. Moreover, Miller et al. [47] recently showed evidence that consistency [9] is a valuable and important component to evaluate the level of trust a human has in an automated decision system. Consistency and reliability are closely related as reliability means repeated and consistent functioning.

In the information technology domain, the most cited trust factors for information systems are credibility, security and reliability. The aptitudes to provide credible information and to convince the user of this credibility are major topics of research. In our reliability definition, the capability to represent data credibility was identified as one aspect of a data model's reliability.

We defined data model reliability as the models ability to preserve and represent data quality. Since data plays a central role in the trust a user places in a database system, reliability of the model is crucial. For that reason, combined with the other evidence provided above, reliability appears to be the most important component.

Understandability is widely recognized as a very important factor for building trust in automation. It is also well known that users must understand the functionality of the automation, and its limitations. However, a processing component is very different from a data model. In the first case, the focus is on the functioning of the process and any associated limitations: What does it do and to what extent? In the latter case, the focus is on a structure, and data organization: What does it represent and how?

For a data model, usefulness and understandability are related. It is almost impossible to find a data model useful if it cannot be understood. On the other hand, it is possible to understand the structure of a database and decide that it is not useful, i.e. that it would be difficult to maintain or reuse. For that reason, it seems reasonable to place understanding before usefulness.

Familiarity definitively helps understanding a data model, but this is not sufficient. For that reason, familiarity comes after understandability and thus usefulness.

Finally, although in the data collection process reputation is important, reputation of the data model maker is less important. See 6.10, for more details about the role of reputation in building trust in a data model.

However, it is not clear which one, familiarity or reputation, is more important for trust. If

9. Automation ability to be consistent in performance, ability to produce similar outcomes for identical tasks.

the data model maker's reputation is unknown, familiarity will become more important to build trust. On the other hand, if the user is aware of the designer's reputation, familiarity may become less important in terms of trust. In addition, both are perceived instantaneously because they are directly related to the user's past experience. As the user gains experience with the model, these two components become less important. For all of these reasons, familiarity and reputation are placed on the same level.

This exercise leads to the following ordered classification of trust components to assess a data model:

1. Reliability
2. Understandability
3. Usefulness
4. Familiarity and Reputation.

Further investigations, in particular empirical ones, would be required to validate that these trust components are significantly important (with that order) for users (both operators and scientists) when assessing their trust in a data model.

7.2 Data Model Influence on Trust in a Database System

It was shown that it is theoretically possible to trust a data model. Now the natural question is: does it influence the overall level of trust in a database system?

To tentatively answer this question, we need to go back to the observation made in section 4:

$$\text{Trust in data} \not\Rightarrow \text{Trust in database}$$

Trust in the data is not sufficient to produce trust in the database system containing that data. Data delivery, processing components, data collection, data model and user's context need to be considered when it comes to trust in a database system (see Figure 5). The relative importance of these components is difficult to assess: it depends on the type of user (including their background, self-confidence and the task at hand) and the importance each component plays in the whole system. In any case, when it comes to trusting a database system, there is no doubt that data plays a central role.

It would be wrong to say that the data model has no influence on trust in a database system. To show this, let us consider the opposite proposition: *data model does not influence the level of trust in a database system* and use the scenario presented in section 5.

Suppose a database was developed to persist the information collected by a highly trusted data collection and that the information is required to perform MSA specific tasks. Moreover, let us assume that the data delivery process is trustworthy.

Now suppose that the data model is designed such that initial data quality will deteriorate over time, or that it is difficult to maintain, too complex, or containing non-familiar vocabulary. Based on the arguments presented in this document, the data model will not be trusted by users. However, can someone trust the database built upon this model to perform its activities? It seems reasonable to say that trust in this database system will be negatively affected, even when the database is filled with pertinent and high quality data.

This example illustrates that a data model does play a role in the level of trust a user places in a database system. It is therefore important to bring a nuance on that affirmation. As mentioned before, without trust in data there cannot be trust in the database system containing it. Therefore, a data model can be viewed as a multiplication factor: it can increase or decrease the level of trust in the database. If there is no trust in the data, no data model, as good and trustworthy as it can be, can make the whole database system trustworthy.

7.3 Hypothesis Validity

Considering the theoretical arguments presented in this document, the initial hypothesis appears valid. When standards and best practices are applied to the data modelling exercise, elements of reliability, familiarity, understandability, usefulness and reputation appear to be satisfied. Since these elements are also trust components, it is reasonable to conclude that a database system built upon a data model utilizing standards and data modelling best practices would be more trusted by the community utilizing the data contained within the database.

8 Use-Case with the National Information Exchange Model for the Maritime Domain

This section analyzes the NIEM-Maritime data model from a trust perspective, using the pertinent trust components to evaluate the model. Being an international maritime data model built to facilitate data exchange, the NIEM-Maritime follows best practices and standard design principles. The purpose of this exercise is to assess the data model in terms of the trust components identified as being important to data modelling. If NIEM-Maritime is found to score high on the trust components, then this would mean that using that particular data model to build a database would increase the level of trust in the database.

Section 8.1 provides the NIEM-Maritime history and scope. Sections 8.2 to 8.6 evaluate NIEM-Maritime with the trust components and section 8.7 concludes the exercise.

8.1 Background

NIEM-Maritime is the result of several efforts to create a common data model for maritime information exchange across agencies. The NIEM-Maritime is an integration of MIEM, as the maritime extension to NIEM, and Universal Core data model (UCore) products and services.

Subsections 8.1.1 to 8.1.3 describe each of these initiatives while 8.1.4 focuses on NIEM-Maritime technical aspects in order to scope the current trust assessment exercise.

8.1.1 MIEM

The MIEM was created in 2006 as part of the Comprehensive Maritime Awareness (CMA) Joint Capability Technology Demonstration [10] sponsored by the US Navy and the Department of Defense. MIEM provides a semantic model, embodied in an eXtensible Markup Language (XML) schema, for tracking people, cargo, vessels, and facilities, as well as relationships among them including threats, anomalies, and other events. The CMA developed the model in collaboration with the Maritime Domain Awareness (MDA) community of interest. As a result, the MIEM meets maritime specific information sharing requirements identified by the MDA inter-agency Data Sharing Community of Interest (COI).

10. The vision is to share maritime shipping information and tracks throughout the world to deter use of commercial maritime shipping for terrorism, weapons proliferation, drugs, piracy, and human trafficking [48].

8.1.2 NIEM

The NIEM [11] represents a collaborative partnership of US agencies and organizations across all levels of government (federal, state, tribal, and local) and with the private sector. It is designed to develop, disseminate and support enterprise-wide information exchange standards and processes that can enable jurisdictions to effectively share critical information in emergency situations, as well as support the day-to-day operations of agencies throughout the nation.

The NIEM reference model includes two categories of reusable components: core components and domain-specific components. These components are illustrated in Figure 8 which can be found in [49]. The core component is in the middle with domain-specific components surrounding it. The core component may be considered a universal or common component. Domain-specific components are understood and managed by a specific COI. Domain-specific components can extend core components and must conform to the NIEM naming and design rules. Domains include a cohesive group of subject matter-experts to ensure some level of authority within the domains they represent, and participate in the processes related to harmonizing conflicts and resolving data component ambiguities. The NIEM version 1.0 was released in 2006.

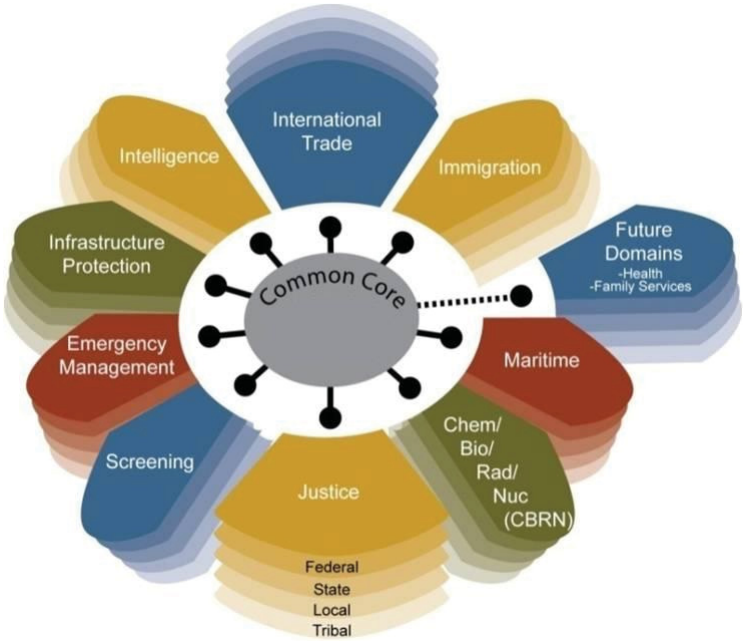

Figure 8: *NIEM Core and Domains components.*

The Department of the Navy and the DoD Executive Agent for MDA worked with the NIEM Program to transition the MIEM into NIEM as its *Maritime* domain component

11. http://www.niem.gov/

(in Figure 8). The resulting version 2.1 of the augmented NIEM including the maritime component was released in September 2009. That is why MIEM is now referred as NIEM-Maritime or simply NIEM. For this document, NIEM-Maritime is preferred.

Recently, the law enforcement and public safety sector in Canada has shown interest in adopting NIEM [50]. Proof of concept investigations are currently being performed by the law enforcement and public safety sectors [51]. Results are recommending NIEM adoption and suggest promotion of NIEM across broader public safety communities [52].

8.1.3 UCore

UCore is a US government project to facilitate sharing of intelligence and related digital content across US government systems. It consists of a vocabulary of commonly exchanged concepts, XML representation of the concepts, extension rules to allow tailoring to specific mission areas, security marking to permit controlled access, and a messaging framework to package and unpackage the content consistently.

UCore version 1.0 was released in 2007 and its focus was to share information between DoD and the Intelligence Community. Version 2.0, released in 2008, was expanded to include the Department of Homeland Security and the Department of Justice. Consequently, it has been designed to be interoperable with NIEM. In addition, the NIEM program has committed to ensuring that NIEM will be compatible with UCore.

8.1.4 NIEM-Maritime

The relationships between the maritime domain, NIEM, and UCore models are illustrated in Figure 9 from [53]. The MDA vocabulary is an integration of MIEM, as the maritime extension to NIEM, and UCore products and services. Each of these reference models started independently but they are now aligning as complementary initiatives with complementary models.

8.1.4.1 Schema

NIEM-Maritime is realized as a XML Schema Description (XSD) file [12]. XSD is a language for constructing data structure specifications, or schemas. XSD is recommended by W3C for describing the rules to which an XML document must conform in order to be considered valid according to that schema. XSD is also used as a formal definition of Web Services Description Language (WSDL) grammar used to describe SOAP-type web services.

12. The NIEM-maritime XSD file can be downloaded from `http://niem.gov/niem/domains/maritime/2.1/maritime.xsd`.

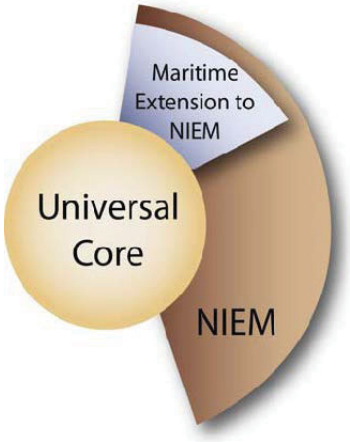

Figure 9: *Relation between NIEM, UCore and Maritime domain.*

XSD files are mainly composed of element and attribute declarations and complex and simple type definitions. The XSD data model includes:

– the vocabulary including element and attribute names, and possibly controlled vocabularies for content (see Isenor and Spears [54])
– the content model (relationships, structure and documentation)
– the data types.

For example, Figure 10 is a graphical representation of an excerpt of the NIEM-Maritime schema. The element *Anomaly* can take values of complex type *AnomalyType*, which is composed of 5 elements of different types (simple and complex).

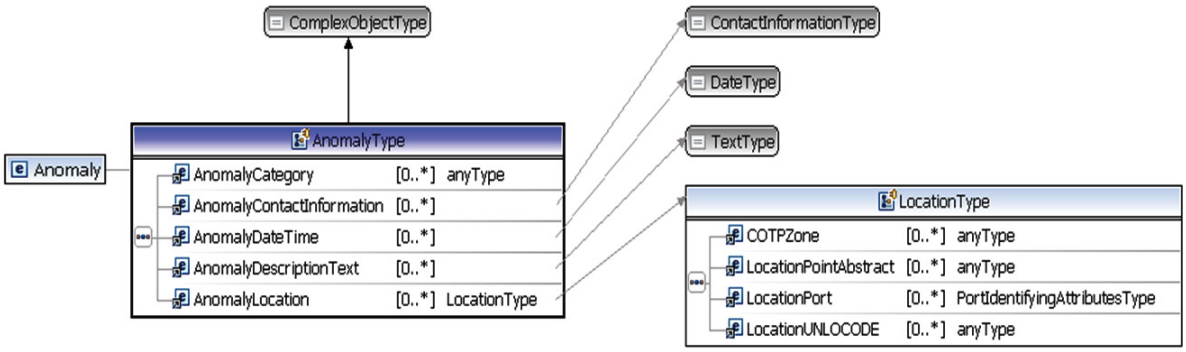

Figure 10: *Anomaly element and data types in NIEM-Maritime.*

8.1.4.2 Purpose

NIEM-Maritime is a data model for maritime information exchange. It is important to note that the purpose of it is to *exchange* maritime information between agencies (e.g., departments, enterprises, etc.). Agencies do not need to operate the same information systems or adopt the same standards and coding schemes for data collection or persisting. Nevertheless, there must be common understanding and semantic consistency in the structure of the data that crosses agency lines if it is to be successfully utilized by the receiving agency.

As noted previously, NIEM-Maritime exists as an XSD file and is therefore not technically a database nor a relational model. Agencies might elect to incorporate NIEM data definitions into new database designs, but this is not required for NIEM conformance. This distinction has to be clearly established, since the goal of this study is to understand the impact of the data model on the level of trust in a *database*.

8.1.4.3 Data Persistence and Database Architecture

For someone interested in persisting maritime data in a database designed with NIEM-Maritime, there are two choices: create a database by mapping the XSD to a relational model or use native XML databases.

It is worth mentioning that creating a relational model from a hierarchical model like NIEM-Maritime is not a straightforward operation. We can observe the same kind of mismatch that may appear between object-oriented programming and a relational database model [55]. That being said, it is technically possible to create a relational model from an XSD file. Tools such as Altova XMLSpy [13] allow the generation of relational database structures from XSD, connection to a relational database, generation of an XSD based on a relational database, and the import and export of data based on database structures. XMLSpy also offers support for NIEM schemas. In any case, the NIEM-Maritime vocabulary, relationships, structure, documentation and types would be preserved, if possible, in the resulting database design.

A native XML DB is a non-relational database storing XML documents in an optimized way and provides XQuery technology [14] and a thin layer of document repository functionality. This kind of database avoids the burden of converting data to XML and vice-versa for export and persistence. Therefore, this type of DB is a good option for persisting standards-compliant XML data, such as NIEM. Consult [56] for a list of XML DB vendors.

The answer to the question *which one (relational or XML DB) is better* depends on the use context. A native XML database is less common than a classical relational database,

13. `http://www.altova.com/xmlspy.html`

14. The XQuery API for Java is currently under development under the Java Community Process. It allows a Java program to connect to XML data sources, prepare and issue XQueries, and process the results as XML.

and using it may have an impact on trust. Scientists are more susceptible to impact by this choice as compared to the operators, especially those scientists who are reluctant to adopt new technologies. Maintenance may require extra training but could be an investment on the medium to long term.

Note that there is also a third option, which can be seen as a compromise between the two options described above. The third option is an XML-aware classic database. An XML-aware classic database maps all XML to a traditional database (such as a relational database), accepting XML as input and rendering XML as output. It allows storing an XML document, such as a NIEM-Maritime compatible XML file, as a data item in a database. Some vendors, such as Microsoft SQL Server [15], also provide XQuery support to query and validate these XML data items. This third option is not considered in the context of this study because it does not imply a database design based on the NIEM-Maritime (it just allows the persistence of NIEM-Maritime XML documents).

8.2 Reliability

Data model reliability is defined as its capability to preserve and represent data quality. It was determined (see Section 6.2.1) that five data quality attributes can be preserved and reflected by a reliable data model. The following subsections analyze the NIEM-Maritime data model reliability in terms of these data quality attributes.

8.2.1 Data Completeness

The migration of the MIEM to NIEM involved a team of 30 platform operators (among others) that initiated a series of 90-day tests. The tests were derived from use-cases ranging from position reporting to suspicious activity reporting, including those supporting DODs JC3IEDM data exchange, as well as MIEM (which has now become NIEM-Maritime). Consequently, the resulting data model was defined within the scope of the use-cases domain which may not include all possible situations. In addition, the NIEM-Maritime data model is still very young and its evolution is based on a spiral approach.

The data model seems to be very detailed in its sub-domains (e.g., anomalies, cargo, customs, etc.). However, some MDA-related concepts such as vessel sensors, vessel weapons, communication frequencies, offshore drilling platform, etc., could not be found in the data model. It is possible these concepts are hidden in other parts of NIEM or NIEM-Maritime remote standards. For example, Automatic Identification System (AIS) reports, which are known to be part of the MIEM data model, could not be found in the NIEM-Maritime data model nor NIEM. It is still unclear if the migration from MIEM to NIEM was incomplete or if the absence of the AIS reports was a design decision.

15. http://technet.microsoft.com/en-us/sqlserver

Another example is related to emergency management. Although there is an Emergency Management module in NIEM 2.1, it is not clear if that domain has been adapted to include maritime emergencies (e.g., a tsunami).

In this case, we view NIEM-Maritime as being complete or incomplete depending on the MSA context.

8.2.2 Data Integrity

Data integrity is directly related to the accuracy of the mapping between data and the data model. If the two are slightly different at a semantic level, it could jeopardize data integrity in the long run. This could be an issue if, for instance, the data model duplicates data in different elements. However, such duplication has not been noted in the case of NIEM-Maritime.

The extent to which a data model preserves data integrity depends also on the DBMS and thus the database implementation (relational or native XML). In the case where a relational database is built using the NIEM-Maritime, the resulting model should be normalized. Since there are few relationships between elements (see sub-section 8.4.1 for details about this aspect), normalization should not be a challenge.

8.2.3 Data Credibility

The extent to which a data model makes data credibility easily perceivable depends mostly on the metadata associated with the data, such as provenance and data quality metrics. NIEM has fairly strong metadata support in its data model, it is reasonable to say that if used properly, the quality of the data will be readily available to the user via related metadata content. For example, the data model supports geographical position quality.

Here are some concrete examples of metadata that can be used to assess data quality and provenance:

***QualityValidityCategoryCodeType*:** A data type for categories that describe the level of probability that the data is trustworthy.

***DateAccuracyIndicatorCodeType*:** A data type for a subjective assessment that indicates belief that date content is exact or accurate.

***MeasureEstimatedIndicator*:** True if a measurement has been estimated or guessed; false otherwise.

The use of such metadata to display provenance and indicate data quality has a positive effect on trust.

8.2.4 Data Representational Consistency

This is the extent to which data are presented in the same format and are compatible with previous data. Since NIEM-Maritime was developed with a team of operators, we can assume that data representation is consistent in format and with data previously utilized by operators.

However, it is not clear to what extent NIEM-Maritime representation is compatible with the Canadian standard representations and formats (e.g., units, processes, hierarchy structure, standard differences, etc.). This aspect could have an impact on trust for non-US operators.

8.2.5 Data Timeliness

Time information is primary and is expected, if not as a typical data entry, it should appear as metadata. The NIEM data model contains [16] all the fields necessary for describing the metadata timeliness associated with a data entry or a data source.

8.3 Familiarity

The familiarity evaluation (see Section 6.4) is based on the use of a vocabulary common to the user community. The fact that MIEM was developed with a team of operators, suggests a certain level of familiarity. However, it is well known that no solution can please everyone. Although including several end-users in the data model development is a recommended strategy to build a trustworthy data model (see guidelines in section 9 for more details), consensus may not be easy to reach when it comes to familiarity. Indeed, familiarity is related to the user's background (e.g., education, training, nationality, etc.), which may differ in a group and across organizations and countries.

In this case, familiarity is partially based on the naming convention used for the schema, which is moderately technical. In addition, we can assume that MDA scientists and operators would be quite familiar with the terms used in the data model.

8.4 Understandability

The evaluation of understandability is based on the evaluation presented in Section 6.5.

NIEM-Maritime, and more generally, the overall NIEM structure, is complex. However, the naming convention used does help with understanding. Complexity issues have been

16. The metadata section is located outside NIEM-Maritime but within NIEM

identified and since the model is relatively young, improvements are expected. Documentation could be improved and as usual with XML-based files, good applications are required to explore and edit the model.

NIEM-Maritime may represent a challenge to learn, especially for users not familiar with XML, the structure, and naming convention. However, since there are few elements in the model, the understandability issue is less problematic. In addition, constant efforts are made to improve documentation that may eventually compensate for the structural complexity.

Overall, understandability is partially mitigated by the naming convention but is still an issue because of the *vertical* nature of the NIEM model.

8.4.1 Structure

The NIEM data model is closer to a *vertical/linear* model as compared to *tree-like/2D* model. There are few relations between elements. Consequently, accessing a piece of data is quite an efficient process. However, the result is that all elements appear be on the same level and relations between the elements is not obvious. This flattening effect makes the model less understandable.

The naming convention minimizes this impact by providing an efficient way to describe elements. Standard use of naming conventions allows users to easily understand and follow the meaning of NIEM elements while defining the local elements.

Moreover, according to findings of the NIEM focus group meeting held in 2008 [57], NIEM is still too complex, this impacting negatively on user understanding. There are too many options/choices (i.e., data elements, etc.). Complexity allows for flexibility but creates a barrier to adoption and use. In addition, a need to have the capability to superset certain subsets that have been created was identified. Since the referred document was produced in 2008, prior to the addition of the maritime domain, it is not clear to what extent it applies to NIEM-Maritime.

8.4.2 Documentation

Documentation is minimal but what exists is easily accessible [17] and includes an advanced search capability. Nevertheless, it may be difficult to find some elements when the exact element name is not known in advance even using the search capability.

Again according to observations of the NIEM focus group meeting held in 2008 [57], there is a need for improved description of NIEM components. It was identified that contextual

17. Documentation about the NIEM-Maritime can be found at `http://www.schemacentral.com/sc/niem21/s-maritime.xsd.html`.

definitions, extra information about specific use, and more complete definitions, would be helpful. This is especially true for these elements where present definitions match the element names.

Note that there is considerable documentation about NIEM (e.g., tutorial, use case, naming convention guide, etc.) available at the NIEM website, and documentation is expected to improve in the future.

Moreover, since the NIEM-Maritime is physically represented as an XSD file, there is a requirement to use XML visualization tools to explore it, and XSD editors to extend and modify it (e.g., XMLSpy, Eclipse model Development Tool, etc.). Such visualization tools are available on line at the NIEM web site [18].

8.5 Usefulness

The usefulness evaluation is based on Section 6.7.

NIEM (and NIEM-Maritime) was developed to effectively and efficiently share critical information across agencies. Data model usefulness is therefore a central concern. However, NIEM was developed to exchange information, not persist the information. Data model usefulness requirements may differ depending on the context being either information exchange or persistence.

Assuming the data contained in a database is useful for a user, its structure can be considered useful or not. Usefulness can be broken down into lower level components to allow measurement for the context of the data model. The following subsections analyze the components of usefulness for the NIEM-Maritime data model, in the context of data persistence, using the components of Section 6.7.

8.5.1 Relevance

This data model was developed in close collaboration with the MDA community. Putting aside some possible incomplete aspects of the data model, all the parts of the data model core are relevant to the end-user. This is not surprising as the first data model iteration, which provided the basic core and the scope (custom, cargo, terrorism, justice, etc), was constructed from use-cases identified in the MDA community. It is understood that the NIEM will follow a spiral process where the functionalities will be reevaluated and corrected (or added) in the next iterations.

We underline again the fact that NIEM was designed for data exchange and it is understood that the scientist is responsible for rearranging the transferred data into a database system relevant to their application.

18. `http://tools.niem.gov/niemtools/home.iepd`.

Relevance may not always be a *transferable* concept, i.e., what is relevant to certain users in a given context may not be relevant for other users. Some work has to be performed to make sure that the NIEM-Maritime is relevant for the database system users. However, chances are good that this model is relevant for a majority of users in a MDA context.

In conclusion, the NIEM data model is usually relevant (within a domain) but it may require some work/processing from the user in a context other than information exchange.

8.5.2 Reusability

As mentioned before, NIEM-Maritime is closer to a vertical/linear model than a typical tree-like/2D model. This vertical model has only a few links with other parts of the model making the elements more independent and therefore, more reusable. The naming convention provides a framework for defining elements, from generic to very precise, thus providing a spectrum of granularities offering multiple reuse opportunities.

As part of the NIEM structure, the NIEM-Maritime module facilitates sharing information between different levels of government, which is useful for information sharing related to operations such as fighting terrorism and drug trafficking. This is not surprising as the main purpose of the NIEM schema is based on reusability and data exchange. The schema can be reused in different contexts, assuming the completeness problems mentioned above (section 8.2.1) do not interfere with the new implementation or could at least be solved by using data from other NIEM modules.

Overall, NIEM-Maritime satisfies the reusability criteria.

8.5.3 Flexibility

Data model flexibility is defined as the model's ability to evolve. The following NIEM-Maritime characteristics make it a flexible data model because they limit the impact of the modifications resulting from this evolution:
– Few relations between elements.
– *Vertical/linear* structure.
– Great number of data type options.

On the other hand, the cost of change may vary depending on the usage of the data model (e.g., data exchange, database system for large or restricted group of users, etc.). As NIEM (and NIEM-Maritime) is part of a largely used and working standard, the cost could be quite high in terms of organizing and conducting meetings aimed at following the authorization process for the modifications.

In addition, one has to consider two different scenarios:

– Addition of new elements into the data model: may have low impact on databases and be backward compatible.
– Moving things around in the data model: may result in drastic changes in the database and jeopardize backward compatibility.

If backward compatibility is not an issue, then it would be safe to say that NIEM-Maritime is quite flexible. If not, flexibility would be reduced to the ability of the model to easily add new features.

8.5.4 Cost of Knowledge Acquisition

One of the important aspects of knowledge acquisition is the ratio of retrieval time to efforts needed for extracting a piece of information. If the data model is such that it takes prolonged time and extra effort to get important information, then the model is considered not useful. The cost of information acquisition can also be aggravated by a non-efficient hardware and/or software (Operating System/DBMS).

The cost of knowledge acquisition in this particular case is low due primarily to the fact that the model structure is close to a vertical/linear structure. Consequently, accessing a piece of data is quite efficient.

As a result, the low cost of knowledge acquisition for NIEM-Maritime makes it useful, which has a positive impact on trust.

8.6 Reputation

The reputation evaluation is based on Section 6.10.

The NIEM data model benefits from large recognition in the US, which is spreading to Canada as well. This reputation is the result of a large multi-disciplinary consultation process which started with a successful core set.

MIEM was first developed by Rick Hayes-Roth [19] and his team. Their work has been recognized and resulted in the integration of the MIEM as the authoritative MDA model for the NIEM, being adopted by both the Navy and Department of Homeland Security (DHS).

This good reputation is also based on the following facts:
– the large and qualified community involvement in the development of NIEM,
– the successful implementation of the data model in a variety of real life applications,
– the fact that the same model can be used in various government departments.

19. http://en.wikipedia.org/wiki/Rick_Hayes-Roth

However, one has to consider the relatively young age of the model, which still has to go through the test of time and implementation. As stated before, reputation has more importance when experience with the data model is still in its infancy.

Therefore, the NIEM data model benefits from an excellent reputation. In Canada, Public Safety Canada has developed an increasing interest in NIEM. It is thus safe to assume that NIEM-Maritime reputation would have a positive impact on trust.

8.7 Concluding Remarks Regarding NIEM

Assessment of the NIEM-Maritime data model is summarized in Table 1.

This exercise allows us to conclude that NIEM-Maritime is a data model likely to instill trust. In addition, from the conclusions drawn in this study, using this data model could influence positively the level of trust in databases built from it.

However, some aspects of NIEM-Maritime and its use could negatively impact trust in a database system. The impact depends on the context in which the database system is used:
- The fact that it is designed for data exchange. It is the database development team's responsibility to create a proper design (which depends also on the type of database) based on NIEM-Maritime.
- Use of metadata: should provide the user with the feeling that the data model preserves data quality.
- Complexity, which impacts understandability and database maintenance.
- Stage of development: NIEM-Maritime domain is at an early stage of deployment and is expected to evolve in the future.

Other factors concerning the resulting database system should also be considered: type of users, data collection process, data quality, data delivery, processing, and information presentation.

This exercise shows that using a good data model standard is a safe way to instill trust, but other factors must be considered. It is not recommended to only rely on the standard reputation. Factors identified above are to be considered in order to produce a trustworthy data model.

To conclude, it is worth mentioning again that data plays the central role in the trust a user places in a database system. We based our NIEM-Maritime analysis solely on the model, assuming that the data would be of high quality.

Trust Components	Sub-Components	Assessment
Reliability	Data Completeness	Poor to Good; depends on the MSA activity
	Data Integrity	Good; this can be impacted by the database architecture
	Data Credibility	Good; if metadata is used properly
	Data Representational Consistency	Good inside US (TBD for Canada)
	Data Timeliness	Good; if metadata is used properly
Understandability	Structure	Weak
	Documentation	Could be improved
Usefulness	Relevance	Very Good
	Reusability	Good
	Flexibility	Good
	Cost of Knowledge Acquisition	Excellent
Familiarity		Excellent
Reputation		Excellent

Table 1: *Trust Assessment for NIEM-Maritime. Ordering of the trust components is according to relative importance of the components as identified in Section 7.1. Assessment is made on a scale consisting of Very Poor; Poor; Good; Very Good; Excellent.*

9 Concluding Remarks

This report has dealt with the sub-components of trust, in a thought analysis that assessed how a data model may be related to these sub-components. The analysis indicates that reliability, understandability, usefulness, familiarity and reputation are the sub-components that capture the concept of trust in a data model.

As a means to conclude the analysis, we now link specific data modelling practices and methods to the trust components. Section 8 provided an assessment of the NIEM data model in terms of the five trust sub-components and highlighted how the NIEM supports the specific sub-component, through design practices or implementation. Using these design practices, we now construct more general statements about what data modelling practices contribute to the five trust sub-components. This results in general guidelines on how to construct a data model while taking into account the five trust sub-components. These guidelines build on the findings from section 8.

Each guideline description is structured as follows:

1. overview of the guideline;
2. trust components covered by the application of the guideline;
3. trade-offs involved in the guideline application.

The relationship between the data modelling practice and the trust sub-component are summarized in Figure 11.

9.1 Apply Data Modelling Standards

As mentioned in section 3.3.2, the types of standards relevant to designing data models are: standard context, analysis standards, naming standards and standard attribute formats. Prior to developing the data model, an investigation of all of these types of standards is recommended.

Standard context represent one of the more contentious choices. Communities are often divided regarding which model is the best. From a trust perspective, it is recommended to select a model based on the user's preference. Scientist users are usually more demanding regarding database structure. Indeed, part of their job is to maintain and build applications that utilize the database system. In addition, as seen in section 6, scientific tasks are linked to more data model requirements that are related to trust components (flexibility, re-usability, etc.).

In any case, context standards are a priority (over other standards) in terms of trust. Context standards will improve the perceived reliability, familiarity, understanding and usefulness. Other standards are more related to reliability. Several studies (see [58] for an overview)

demonstrate the importance of organizational issues in data modelling as opposed to the relative unimportance of technical issues. A parallel can be made between this observation and the importance of context standards relative to other standards. This is why it is also recommended to communicate and discuss about the selected context standards to the development team.

Although the priority should be on context standards, other standards must also be applied. West [14] and [16] provides an overview on the importance of analysis standards.

9.1.1 Related Trust Components

Reliability, Familiarity, Understandability, Usefulness, Reputation.

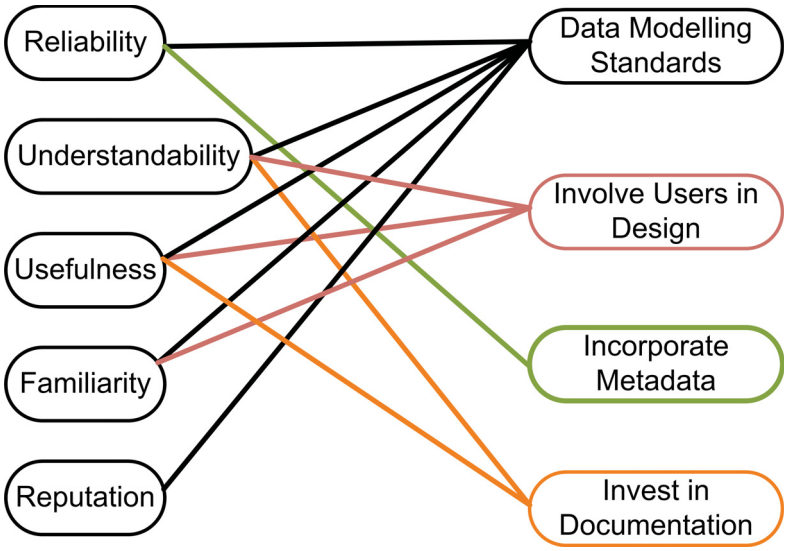

Figure 11: *A color-coded depiction of data modelling practices as related to trust sub-components that have been identified as important for data model trust development. Lines indicate the trust sub-component (left panel) that are influenced by the data modelling practice (right panel).*

9.1.2 Trade-Offs

Time and expertise are the major trade-offs. To do things properly, it takes time and experienced and/or talented developers. Moreover, because there is more than one standard, there is a need to decide which one to use, which requires meetings and consultations.

Depending on the size and scope of the project, insufficient data modelling standards may imply high costs (see Figure 7). These engendered costs usually surpass costs to develop a quality and trustworthy data model. Moreover, even for a small project, applying data

modelling standards can have great benefits. It will be easier to identify potential of re-use of the resulting databases.

9.2 Involve Users in the Design Process

Involving users in the data model design process will help to produce a useful, understandable, and familiar data model. One of the most often-mentioned data warehouse failure factors is the lack of user community participation in the development process (see [59] and [60] for more details).

Including user participation in data modelling means:

1. DB user identification (size of the user group, needs, environments, etc.).
2. Building a competent user/developer team.
3. Scoping the level of user involvement.

Maier [58] proposes an assignment of data modelling responsibility to a selected group of users. This could be used as a starting point to building and managing the user/developer team.

9.2.1 Related Trust Components

Usefulness, Understandability, Familiarity.

9.2.2 Trade-Offs

Building a competent user-developer team is difficult. In particular, choosing the *right* users is critical. Problems with the assignment of responsibilities can occur especially if the user/developer team is composed from different organizational departments or institutions. On this topic (for data warehousing projects), Greenfield [60] notes:

> In many data warehousing projects, it is not uncommon for the information system organization to find one to a handful of users whose ”needs” go way beyond those of most of the data warehouse users. Usually, the need is for a far greater level of detail and/or for far more history and/or for a series of reports of both a high deal of technical and business complexity. It can be quite expensive and time consuming to satisfy the needs of these far more demanding users. On the other hand, these users can have a peculiar need that is especially beneficial to the business and/or can be people whose support is vital to the success of the project.

Another aspect to consider is politics. The more people involved within the organization or community, the more politically sensitive the situation may become (see Demarest [59]).

9.3 Incorporate Metadata

Metadata can be incorporated to help the user in assessing the data quality. At the design step, it is recommended to identify relevant information to assess data credibility and include it in the model. Such information can be quality metrics and/or provenance information. Provenance is particularly interesting since it is objective information that increases the transparency of the process.

9.3.1 Related Trust Components

Reliability.

9.3.2 Trade-Offs

Main trade-offs to incorporate metadata are time and expertise. Moreover, metadata definition requires a good knowledge of the collection process used to populate the database.

9.4 Invest Time in Documentation

As with any kind of software, quality documentation plays an instrumental role in understandability. Moreover, good documentation (see section 6.5.1.2 for details about required data model specific documentation) helps users to identify reuse opportunities.

Note that aesthetics are also important. This means that diagrams must be well balanced, interesting, and uncluttered.

9.4.1 Related Trust Components

Understanding, Usefulness.

9.4.2 Trade-Offs

Again, time is the trade-off to produce quality documentation. That being said, documentation should be planned as part of the development. Documentation is always an investment in the future.

References

[1] Adams, B.D., Bruyn, L.E., Houde, S., and Angelopoulos, P.A. (2003), Trust in Automated Systems: Literature Review, (DRDC Toronto CR 2003-096) Defence R&D Canada – Toronto.

[2] Artz, D. and Gil, Y. (2007), A survey of Trust in Computer Science and the Semantic Web, *Web Semantics: Science, Services and Agents on the World Wide Web*, 5, 58–71.

[3] Isenor, Anthony W., Lapinski, Anna-Liesa S., and MacInnis, Andrew (2012), Trust in automated maritime situational awareness systems with application to AIS, (DRDC Atlantic TM 2011-279, in preparation) Defence R&D Canada – Atlantic.

[4] Kelly, C., Boardman, M., Goillau, P., and Jeannot, E. (2003), Guidelines for Trust in Future ATM Systems: A Literature Review, Technical Report European Organisation for the Safety of Air Navigation.

[5] Sheridan, Thomas B. (1992), Telerobotics, Automation, and Human Supervisory Control, Cambridge, MA, USA: MIT Press.

[6] Sheridan, T. B. (1988), Trustworthiness of command and control systems, In *Proceeding of Analysis Design and Evaluation of Man-Machine Systems*.

[7] Dutton, Philip D. (2000), Trust: Issues in the Design and Development of Electronic Commerce Systems, Technical Report Griffith University, School of Computing and Information Technology.

[8] Madsen, Maria and Gregor, Shirley (2000), Measuring human-computer trust, In *Proceedings of the 11 th Australasian Conference on Information Systems*, pp. 6–8.

[9] Park, Eui, Jenkins, Quaneisha, and Jiang, Xiaochun (2008), Measuring Trust of Human Operators in New Generation Rescue Robots, In *Proceedings of the 7th JFPS International Symposium on Fluid Power, TOYAMA 2008*, JFPS '08, pp. 489–492.

[10] Naumann, F. and Roth, M. (2004), Information Quality: How Good Are Off-The-Shelf DBMS, In *Proceedings of the International Conference on Information Quality (ICIQ)*, pp. 260–274.

[11] Simsion, G. (2007), Data Modeling Theory and Practice, USA: Technics Publications, LLC.

[12] Isenor, A.W. and Spears, T.W. (2009), Utilizing Arc Marine Concepts for Designing a Geospatially Enabled Database to Support Rapid Environmental Assessment, (DRDC Atlantic TM 2009-061) Defence R&D Canada – Atlantic.

[13] (2011), Data model (online), Wikipedia, `http://en.wikipedia.org/wiki/Data_model` (Access Date: 2011).

[14] West, M. (1996), Developing High Quality Data Models, (Technical Report 2.1) European Process Industries STEP Technical Liaison Executive.

[15] Levitin, A. and Redman, T. (1995), Quality dimensions of a conceptual view, *Information Processing and Management*, 31, 81–88.

[16] West, Matthew (2011), Developing High Quality Data Models, Burlington, US: Morgan Kaufmann.

[17] Llinas, J., Bisantz, A., Seong, Y., Jian, J., and Drury, C. G. (1998), Studies and analyses of aided adversarial decision-making. Phase 2: Research on Human Trust in Automation, (Technical Report 14260–2050) Center for Multisource Information Fusion, State University of New York at Buffalo.

[18] Perry, W., Signori, D., and Boon, J. (2004), Exploring Information Superiority: A Methodology for Measuring the Quality of Information and Its Impact on Shared Awareness, Technical Report RAND.

[19] Golbeck, Jennifer (2006), Trust on the World Wide Web: A Survey, *Foundations and Trends in Web Science*, 1(2), 131–197.

[20] Fogg, B. J., Soohoo, Cathy, Danielson, David R., Marable, Leslie, Stanford, Julianne, and Tauber, Ellen R. (2003), How do users evaluate the credibility of Web sites?: a study with over 2,500 participants, In *Proceedings of the 2003 conference on Designing for user experiences*, DUX '03, pp. 1–15, New York, NY, USA: ACM.

[21] Grabner-Kräuter, S. and Kaluscha, E. A. (2003), Empirical research in on-line trust: a review and critical assessment, *International Journal of Human-Computer Studies*, 58, 783–812.

[22] Moray, N. and Inagaki, T. (1999), Laboratory studies of trust between humans and machines in automated systems, *Transactions of the Institute of Measurement and Control*, 21(4-5), 203–211.

[23] Wang, Richard Y. and Strong, Diane M. (1996), Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal on Management of Information Systems*, 12, 5–33.

[24] Data Integrity (online), BusinessDictionary.com, `http://www.businessdictionary.com/definition/data-integrity.html` (Access Date: 2011).

[25] Groth, Paul, Jiang, Sheng, Miles, Simon, Munroe, Steve, Tan, Victor, Tsasakou, Sofia, and Moreau, Luc (2005), An Architecture for Provenance Systems, Technical Report University of Southampton.

[26] Bhagwat, Deepavali, Chiticariu, Laura, Tan, Wang-Chiew, and Vijayvargiya, Gaurav (2004), An annotation management system for relational databases, In *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB '04, pp. 900–911, VLDB Endowment.

[27] (2010), W3C Provenance Incubator Group Wiki (online), W3C, `http://www.w3.org/2005/Incubator/prov/wiki/Main_Page` (Access Date: 2011).

[28] Ceolin, Davide, Groth, Paul, and Van Hage, Willem Robert (2010), Calculating the Trust of Event Descriptions using Provenance, In *Proceedings Of The SWPM 2010, Workshop At The 9th International Semantic Web Conference, ISWC-2010*.

[29] McGee, William C. (1976), On user criteria for data model evaluation, *ACM Transactions on Database Systems*, 1, 370–387.

[30] Serrano, Manuel, Trujillo, Juan, Calero, Coral, and Piattini, Mario (2007), Metrics for Data Warehouse Conceptual Models Understandability, *Inf. Softw. Technol.*, 49, 851–870.

[31] Genero, M. and Piattini, M. (2002), Quality in Conceptual Modeling, pp. 13–44, Kluwer Academic Publisher.

[32] Poels, G. and Dedene, G. (2000), Measures for assessing dynamic complexity aspects of object-oriented conceptual schemes, In *Proceedings of the 19th international conference on Conceptual modelling*, ER'00, pp. 499–512, Berlin, Heidelberg: Springer-Verlag.

[33] Calero, C., Piattini, M., and Genero, M. (2001), Metrics for Controlling Database Complexity, pp. 48–68, Hershey, PA, USA.

[34] Moody, Daniel L. (1998), Metrics for Evaluating the Quality of Entity Relationship Models, In *Proceedings of the 17th International Conference on Conceptual Modeling*, ER '98, pp. 211–225, London, UK: Springer-Verlag.

[35] Juhn, S. and Naumann, J. D. (1997), The Effectiveness of Data Representation Characteristics on User Validation, In *Proceedings of the International Conference on Information Systems*, pp. 212–226.

[36] Nordbotten, J. C. and Crosby, M. E. (1999), The effect of graphic style on data model interpretation, *Information Systems Journal*, 9(2), 139–156.

[37] Moody, Daniel L. (2002), Complexity Effects on End User Understanding of Data Models: an Experimental Comparison of Large Data Model Representation Methods, In *XTH European Conference on Information Systems (ECIS 2002)*.

[38] Maier, R. (1996), Benefits and Quality of Data Modelling - Results of an Empirical Analysis, In *Proceedings of the 15th International Conference on Conceptual Modeling*, ER '96, pp. 245–260.

[39] Moody, Daniel L. (1997), A Multi-Level Architecture for Representing Enterprise Data Models, In *Proceedings of the 16th International Conference on Conceptual Modeling*, ER '97, pp. 184–197.

[40] Lucia, De, Gravino, Oliveto, and Tortora (2008), Data Model Comprehension: An Empirical Comparison of ER and UML Class Diagrams, In *Proceedings of the 2008 The 16th IEEE International Conference on Program Comprehension*, ICPC '08, pp. 93–102, Washington, DC, USA: IEEE Computer Society.

[41] Hoxmeier, John A. (2001), Dimensions of Database Quality, pp. 28–47, Hershey, PA, USA.

[42] Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. (1993), The Cost Structure of Sensemaking, In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, CHI '93, pp. 269–276, New York, NY, USA: ACM.

[43] Card, S. K., Pirolli, P., and Mackinlay, J. D. (1994), The cost-of-knowledge characteristic function: display evaluation for direct-walk dynamic information visualizations, In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, CHI '94, pp. 238–244, New York, NY, USA.

[44] Torenvliet, G., Euerby, A., Scott, S., and Histon, J. (2011), Investigating virtual social networking in the context of military interoperability: Year 1 Report, (DRDC Atlantic CR 2010-308) Defence R&D Canada – Atlantic.

[45] Muir, B. (1994), Trust in automation: Part 1. Theoretical issues in the study and human intervention in automated systems, *Ergonomics*, 37(11), 1905–1923.

[46] Muir, B. and Moray, N. (1996), Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation., *Ergonomics*, 39(3), 429–460.

[47] Miller, Janet E and Perkins, Leeann (2010), Development of Metrics for Trust in Automation, *Human Performance*.

[48] Dwyer, Chris (2007), Comprehensive maritime awareness (CMA) joint capabilities technology demonstration (JCTD), *Proceedings of SPIE, the International Society for Optical Engineering*, pp. 65781I.1–65781I.5.

[49] Hoang, Anthony (2009), The National Information Exchange Model, In *Roadmap to Emergency Data Standards Roundtable*, OASIS Emergency Interoperability Summit 2009, Baltimore, US.

[50] PS-CIOD Program (2010), Public Safety Canada (PSC) Interoperability – Data Exchange Standards.

[51] Canada, Public Safety (2011), Communications Interoperability Action Plan for Canada, Technical Report Canada.

[52] Hunter, Linda (2011), Public Safety Canada: Interoperability, In *Developing Enabling Standards for NIEM and ISE Workshop*, OMG Technical Meeting Special Event, Arlington, USA.

[53] Officer, Chief Information (2008), Maritime Domain Awareness Architecture Management Hub Strategy, Technical Report Department of the Navy, USA.

[54] Isenor, Anthony and Spears, Tobias (2007), A Conceptual Model for Service-oriented Discovery of Marine Metadata Descriptions, *CMOS Bulletin*, 35(3), 85–91.

[55] Clifton, Marc and Long, Mark (2009), Introduction to NIEM and IEPDs. Available at `http://www.codeproject.com/KB/recipes/NIEM_IEPD2.aspx?msg=2864405`.

[56] (2011), XML Database (online), Wikipedia, http://en.wikipedia.org/wiki/Xml_database (Access Date: 2011).

[57] Perbix, Mark (2008), NIEM Focus Group Report, Technical Report NIEM.

[58] Maier, Ronald (2001), Organizational Concepts and Measures for the Evaluation of Data Modeling, pp. 1–27, Hershey, PA, USA.

[59] Demarest, Marc (1997), The Politics of Data Warehousing. http://www.noumenal.com/marc/dwpoly.html.

[60] Greenfield, Larry (2000), Data Warehousing Political Issues. http://www.dwinfocenter.org/politics.html.

List of symbols/abbreviations/acronyms/initialisms

AIS	Automatic Identification System
API	Application Programming Interface
CMA	Comprehensive Maritime Awareness
COI	Community of Interest
DB	Database
DBMS	Database Management System
DHS	Department of Homeland Security
DQ	Data Quality
ER	Entity-Relationship
GUI	Graphical User Interface
HTML	HyperText Markup Language
JC3IEDM	Joint Consultation Command and Control Information Exchange Data Model
MDA	Maritime Domain Awareness
MIEM	Maritime Information Exchange Model
MSA	Maritime Situational Awareness
NIEM	National Information Exchange Model
OS	Operating System
OWL	Web Ontology Language
RDBMS	Relational Database Management System
RDF	Resource Description Framework
SOA	Service Oriented Architecture
SOAP	Simple Object Access Protocol
UCore	Universal Core data model
UML	Unified Modelling Language
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
WS	Web Service
WSDL	Web Services Description Language
XML	eXtensible Markup Language
XSD	XML Schema Description

Glossary

consistency

the extent to which the data are presented in the same format and are compatible with previous data.

completeness

the extent to which data are of sufficient breadth, depth and scope for the task at hand.

credibility

the extent to which data are accepted or regarded as true, real, and credible.

data integrity

The accuracy and consistency of stored data, indicated by an absence of any alteration in data between two updates of a data record. Data integrity is imposed within a database at its design stage using standard rules and procedures, and is maintained using error checking and validation routines.

familiarity

the adherence of the data model to the procedures, terms, and cultural norms of the user.

flexibility

the ease with which the data model can cope with business and/or regulatory change.

predictability

the estimation or anticipation of an outcome given a specific and understood set of inputs.

provenance

the process that lead to the creation of that datum.

relevance

the extent to which the data model is applicable and helpful to the task at hand.

reliability

the data models capacity to preserve and represent data quality.

reusability

the extent to which the data model allows data to be used for purposes beyond those for which the data was initially collected.

robustness

the systems ability to perform under a variety of circumstances.

timeliness

the extent to which the age of the data is appropriate for the task at hand.

understandability

how the user perceives their comprehension of the data model.

This page intentionally left blank.

Distribution list

DRDC Atlantic CR 2011-107

Internal distribution

1 F. Desharnais

1 T. Hammond

1 M. Hazen

2 A. W. Isenor (1 hardcopy; 1 CD)

1 L. Lapinski

1 A. MacInnis

1 D. Schaub

1 S. Webb

3 Library (1 hardcopy; 2 CDs)

Total internal copies: 12

External distribution

Department of National Defence

1 Library and Archives Canada,
Atten: Military Archivist
Governments Records Branch

1 NDHQ/DRDKIM 2-2-5

2 Marie-Odette St-Hilaire (1 hardcopy; 1 CD)
OODA Technologies Inc.
4891 Av. Grosvenor,
Montreal Qc, H3W 2M2

1 Laura Ozimek DSTM 5
DRDC Corporate
305 Rideau Street
Ottawa

1 Jack Pagotto
DRDC CSS
222 Nepean St.
Ottawa, Ontario K1A 0K2

1 Stéphane Paradis
DRDC Valcartier
2459 Boul. Pie XI Nord
Québec City, Québec G3J 1X5

1 Jean Roy
DRDC Valcartier
2459 Boul. Pie XI Nord
Québec City, Québec G3J 1X5

1 Barry Walker
CMS DGMFD DMIMR
NDHQ 101 Colonel By Drive
Ottawa ON
K1A 0K2

Total external copies: 9

Total copies: 21

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

OODA Technologies Inc.
4891 Av. Grosvenor, Montreal Qc, H3W 2M2

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED
(NON-CONTROLLED GOODS)

DMC A REVIEW: GCEC JUNE 2010

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

Implicit Trust in a Data Model

4. AUTHORS (Last name, followed by initials – ranks, titles, etc. not to be used.)

St-Hilaire, M.-O.; Mayrand, M.; Isenor, A.W.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information. Include Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
October 2011	86	60

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Contract Report

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

Defence R&D Canada – Atlantic
P.O. Box 1012, Dartmouth, Nova Scotia, Canada B2Y 3Z7

9a. PROJECT NO. (The applicable research and development project number under which the document was written. Please specify whether project or grant.)	9b. GRANT OR CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
Project 11HL, 11HO	W7707-115137-Callup1

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC Atlantic CR 2011-107	

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

(X) Unlimited distribution
( ) Defence departments and defence contractors; further distribution only as approved
( ) Defence departments and Canadian defence contractors; further distribution only as approved
( ) Government departments and agencies; further distribution only as approved
( ) Defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11)) is possible, a wider announcement audience may be selected.)

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

As an assessment of databases used to store Maritime Situational Awareness data, the hypothesis is posed that *a database built upon a data model, utilizing international standards, recognized and accepted data modelling concepts, best practices, etc. would be more trusted by the community utilizing the data contained within the database*. In this work, the validity of this hypothesis was investigated and assessed by decomposing trust into the sub-components: predictability, dependability, faith, reliability, robustness, familiarity, understandability, explication of intention, usefulness, competence, self-confidence, and reputation. An analysis of how these components are expressed in the context of a database system and in particular, how they impact the data model, was performed. The analysis indicates that reliability, understandability, usefulness, familiarity and reputation are the components that capture the concept of trust in a data model. These components were then applied in an analysis of the National Information Exchange Model-Maritime data model, essentially grading the model against the applicable trust components. Results vary from a poor grade on aspects of reliability, to excellent in terms of familiarity and reputation.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

trust
database
data model
maritime situational awareness

This page intentionally left blank.

Defence R&D Canada

Canada's leader in defence
and National Security
Science and Technology

R & D pour la défense Canada

Chef de file au Canada en matière
de science et de technologie pour
la défense et la sécurité nationale

www.drdc-rddc.gc.ca