16th ICCRTS

“Collective C2 in Multinational Civil-Military Operations”

Science and Technology Issues Relating to Data Quality in C2 Systems

Topics:
(4): Information and Knowledge Exploitation

Jonathan Agre
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
USA
+1-703-933-6522
jagre@ida.org

M. S. Vassiliou
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
USA
+1-703-845-4385
mvassili@ida.org

Corinne Kramer
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
USA
+1-703-578-2805
ckramer@ida.org

Point of Contact:

M. S. Vassiliou
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
USA
+1-703-845-4385
mvassili@ida.org

Abstract

A Command and Control (C2) system depends crucially on having high-quality underlying data. There is still no “best” set of data quality dimensions and metrics for C2. We consider various sources of data quality criteria, such as the sixteen data quality dimensions identified by the Total Data Quality Management (TDQM) research community, and the dimensions identified by the ISO 8000 and 25012 standards. We map these dimensions against the criteria commonly applied by the intelligence community (IC) and those identified by various parts of the US DoD, and compare them in terms of relevance to C2. We also describe metrics for these data quality dimensions, and discuss examples of data quality issues in existing C2 systems that have drastic real-world consequences. We examine three important characteristics relevant to C2: interoperability, data volume and trustworthiness, and identify open research areas. We conclude that once an accepted set of data quality characteristics and associated metrics for C2 is available there is a good case for explicitly incorporating it into current and future C2 systems.

1 Introduction

The transition to a net-centric environment and the increasing automation of command and control (C2) functions make the quality of the underlying data upon which decisions and actions are based critical to success. Operating on bad data can have serious consequences, especially in a military context. In the commercial arena, it is estimated that operating on poor data has an economic cost of about $600 B annually [1]. A few of the many side-effects of poor data quality include delays due to reconciling data, loss of credibility, customer dissatisfaction, compliance problems, delivery delays, lost revenue, and loss of trust in the automation and computing systems. Properties that reflect good data- accuracy, integrity, provenance and timeliness, as well as the ability to share the data with others and to have a common understanding of its meaning- are intuitively desirable but are not routinely incorporated in today’s complex systems, in part, because the underlying architectures do not make data quality a primary objective of system design. In the military C2 domain, the effects of poor data can have even more disastrous consequences than in other domains. Making quality considerations an inherent part of the design and maintenance processes of C2 systems should benefit the decision making. We explore some of the associated challenges and issues.

Data is a resource that must be managed, protected and preserved across its life cycle like any other. The dominant issues confronting data management in large enterprises have been frequently reported and include missing or incorrect data, missing or incorrect metadata, redundant data storage, varying data semantics and non-standard data formats. Data portability (freeing the data from stove-piped applications) is also a common concern in both domains. These issues are also of prime concern in C2 systems.

Various investigators have given different definitions of the terms “information” and “data,” depending on the use. For the purposes of this paper, we define the terms as follows:

- **Information** is defined as knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning [2].
- **Data** is defined as the re-interpretable representation of information in a formalized manner suitable for communication, interpretation or processing [2].

In the context of this paper, data includes both raw and processed information and “all data assets such as system files, databases, documents, official electronic records, images, audio files, Web sites, and data access services.” [3].

For C2 functions, data is used to develop situational awareness and a common operating picture by which commanders make decisions and effect control. Commanders require many types of data- ranging from logistics to weather to geospatial to tactical information- to support the various warfighter operations. Data must be collected, analyzed and communicated via various manual and automated messages and exchanged between various C2 systems and people. A commander has little control over the sources that supply data to his C2 systems, especially in times of crisis. Each C2 system may store portions of current data and maintain some amount of past data for historical analysis purposes. The tempo of activity and the volume of data on which the system depends are both rapidly increasing, revealing many stress points in the current systems. In general terms, a modern C2 system is a large, heterogeneous distributed, real-time processing system that is resource limited (bandwidth and computation power) at some of the end-points, with frequent disruptions and highly dynamic information flows. The data is segregated among multiple classification levels and is contained in multiple, distributed storage facilities and heterogeneous databases. As data is delivered with higher frequency and larger volume from more places, decision makers must become more responsive. Modern C2 systems, especially in a coalition environment, are among the most complex systems imaginable.

In this paper we examine a number of important science and technology (S&T) issues relating to data quality in C2 systems. First, we discuss the characterization of the various quality properties of data. We then examine several of these quality characteristics in the context of C2 systems. Finally, we offer some suggestions for further research in several S&T areas to address some of these issues.

2 Data Quality

Data Quality can be simply defined as the fitness for use of the data [4]. A more practical definition is the degree to which data “meets the requirements of its authors, users, and administrators” [5]. The key point to be taken from these definitions is that the generic notion of the quality of data, like many other ideals of quality, is dependent on context, or intended use. Nevertheless, given that data is such a pervasive part of any information technology (IT) system, many ways of partitioning its quality properties have been suggested. In some early data quality research, data was primarily characterized by Accuracy, Completeness, Timeliness, and

Standards (ACTS). This basic list has been expanded over the years in many directions. In particular, since the early 1990s, a Total Data Quality Management (TDQM) research community, [6], has expanded ACTS to sixteen data quality dimensions and successfully used them in assessments of an organization's data quality environment:

- Accessibility - The extent to which data is available or easily and quickly retrievable.
- Amount of Information - The extent to which the volume of data is appropriate for the task at hand.
- Believability - The extent to which data is regarded as true and credible.
- Reputation -- The extent to which information is highly regarded in terms of its source or content.
- Completeness - The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.
- Conciseness - The extent to which data is compactly represented.
- Consistent Representation - The extent to which the data is presented in the same format.
- Ease of Operations - The extent to which data is easy to operate on and applies to different tasks.
- Free-of-Error - The extent to which data is correct and reliable.
- Interpretability - The extent to which data is in appropriate languages, symbols, and units and the definitions are clear.
- Objectivity - The extent to which data is unbiased, unprejudiced, and impartial.
- Relevancy - The extent to which data is applicable and helpful for the task at hand.
- Security - The extent to which access to data is restricted appropriately to maintain its security.
- Timeliness - The extent to which data is sufficiently up-to-date for the task at hand.
- Understandability - The extent to which data is easily comprehended.
- Value Added - The extent to which data is beneficial and provides advantages from its use.

These sixteen characteristics can be grouped into the following four categories [6]: Intrinsic, Accessibility, Contextual and Representational:

- Intrinsic – Accuracy, Reputation, Believability, Objectivity
- Accessibility (Operational) – Accessibility, Access Control
- Contextual – Relevancy, Timeliness, Completeness, Amount of Information, Value Added
- Representational –Conciseness, Consistent Representations, Ease of Operations, Interpretability, Understandability

The intrinsic properties relate to the accuracy and pedigree of the data and do not change depending on environment or intended use. Accessibility, in this usage, refers to the system properties such as how and where the data is stored and the means of protecting the data, such as access control. Contextual properties depend on the application for which the data is used and can have temporal behavior. The representational properties are the more common notions of standardization and interoperability. These categories help to show how the characteristics are related to each other and the environment in which they are situated.

Currently, an International Organization for Standardization (ISO) standard on Data Quality, ISO 8000 [7], is being developed. It is primarily aimed at quality facets of automated information exchange for the purchase of goods. ISO 8000 defines formats for descriptions of master data. It defines data quality using five characteristics: Syntax, Provenance, Completeness, Accuracy and Certification, and considers the processes that are needed to assure data quality. Reference [8] defines *master data* as data held by an organization that describes the independent and fundamental entities for an enterprise. For an organization, this might include descriptions of employees, customers, suppliers, products, services, locations, etc. ISO 8000 Part 110 focuses on requirements for exchange of master data that can be checked through automation [9]. The representation and exchange of information about provenance (Part 120), accuracy (Part 130), and completeness (Part 140) have also been recently published. Provenance information, for example, may include records of creation, extraction, ownership and transfer of ownership of data.

In general, ISO 8000 is oriented towards logistics information, industrial applications or Enterprise Resource Planning (ERP) systems. It has been supported by organizations such as the North Atlantic Treaty Organization (NATO) and the Defense Logistics Information Service (DLIS). DLIS has supported the transition of the Federal Cataloging System and NATO Codification System (NCS) into these open public standards. The Federal Logistics Information System provides automated data on the Federal Catalog System and descriptions of items of supply for the US military using this standard. It serves as the common frame of reference for DoD buyers to communicate with their industrial supplier base [10].

ISO 8000 is closely aligned with several other related data exchange standards such as the ISO 22745 Open Technical Dictionary that defines concepts for describing items, as well as a query interface for accessing the definitions [11]. The Electronic Commerce Code Management Association (ECCMA) Open Technical Dictionary (eOTD) is an ISO-22745 compliant dictionary that has evolved from the NCS and is directed towards the global commercial environment [12]. An eOTD catalog is composed of Extensible Markup Language (XML) files containing information explicitly encoded using ISO-22745 concept identifiers and describes a common supply language for all logistical needs of NATO representing over 31 million reference numbers, 22 million users and 1.5 million organizations.

Another ISO Data Quality standard, ISO/IEC [International Electrotechnical Commission] 25012 ("Data Quality Model") is under development for the domain of software engineering and software quality [13]. This data quality standard is part of a family of standards (25012, 25020, 25021, 25030) defining software system and software engineering quality requirements and measurements, called the SQuaRE standards. The ISO/IEC 25012 document is aimed at structured data stored in computer systems and defines fifteen data quality characteristics divided into two points of view: inherent and external. Inherent data quality is similar to the intrinsic category discussed above and external data quality refers to system-dependent aspects that preserve data quality. All refer to a specific context of use. The fifteen characteristics [13] are:

- Accuracy - The extent to which data has attributes that correctly represent the true value of the intended attribute of a concept or event.
- Completeness - The extent to which subjects associated with an entity have values for all expected attributes and related entity instances.
- Consistency - The extent to which data has attributes that are free from contradiction and coherent with other data.
- Credibility - The extent to which data has attributes that are regarded as true and believable by users.
- Currentness - The extent to which data has attributes that are of the right age.
- Accessibility - The extent to which data has attributes that enable it to be reached, particularly by people who need supporting technology or special configuration because of some disability.
- Compliance - The extent to which data has attributes that adhere to standards, conventions, or regulations in force and similar rules relating to data .
- Confidentiality - The extent to which data has attributes that ensure that it is accessed and interpreted only by authorized users.
- Performance - The extent to which data has attributes that can be processed and provide the expected level of performance by using the appropriate amounts and types of resources under stated conditions.
- Precision - The extent to which data has attributes that are exact or that provide discrimination.
- Traceability - The extent to which data has attributes that provide an audit trail of accesses to the data and of any changes made to the data.

- Understandability - The extent to which data (and associated metadata) has attributes that enable it to be read and easily interpreted by users, and are expressed in appropriate languages, symbols and units.
- Availability – The extent to which data has attributes that enable it to be retrieved.
- Portability – The extent to which data has attributes that enable it to be moved from one platform to another preserving the existing quality.
- Recoverability - The extent to which data has attributes that enable the data to maintain and preserve a specified level of operations and quality, even under failure.

A key difference from ISO 8000 is the exclusion of provenance. There is clear overlap with the TDQM characteristics, but also some key differences, primarily more inclusion from the operational viewpoint. Some of the operational characteristics that are not stressed in TDQM include Performance, Portability, Recoverability and Availability. In ISO/IEC 25012, Compliance refers to adherence to standards and regulations, something TDQM does not explicitly consider. ISO/IEC 25012 also groups the characteristics according to whether they refer to inherent or external data quality characteristics, or both. *Accuracy* through *Understandability* are inherent, *Accessibility* through *Recoverability* are external and *Accessibility* through *Understandability* are both.

The intelligence community (IC) has traditionally been very concerned about data quality. The Joint Military Intelligence Committee identified six characteristics of data quality [14]:

- *Accuracy*: Data and its sources are evaluated for technical errors, misperceptions, or deliberate efforts to mislead.
- *Objectivity*: The data is examined for deliberate distortions and manipulations due to self-interest.
- *Usability*: Data is compatible with a customer’s capabilities for receiving, manipulating, protecting and storing the product and is ready when needed.
- *Relevance*: Information is applicable to customer requirements.
- *Readiness*: Data systems must be responsive to the dynamic requirements of customers.
- *Timeliness*: Data must be available and acted upon when it is required.

These properties have been contrasted with and extended to the sixteen TDQM quality dimensions as described in Reference [15]. These six basic categories above are naturally slanted towards the specific needs of the IC rather than information automation, but since C2 systems rely heavily on intelligence products, their data quality needs clearly overlap.

Data sharing and accessibility are areas that have received much public attention since 9/11. The IC is also very worried about spoofing or the injection of false data that can corrupt decisions or analyses. There is a great need for provenance information to track sources, and the intermediate handling of data to detect deliberate deception attempts. Another concern is that of inconsistent data that can arise from multiple observers. Non-authoritative sources of data are also a persistent problem, and proper weighting is needed. In some C2 systems, such as the Global Command and Control System (GCCS), the data is generally vetted and considered authoritative, while in others, such as the Tactical Ground Reporting (TIGR) System, the data can be entered by any user that observes an interesting event. Both types of systems have their uses, but the differences show that the pedigree of data should be an explicit factor. Another interesting IC and C2 issue is that information that was presented as true may later be found to be untrue, and that this meta-information needs to be disseminated as well. Some data quality properties, such as timeliness and accuracy, can have a more severe impact in a C2 tactical situation. It is not acceptable, for example, to target the wrong building due to incorrect data.

The DoD has recognized data quality as an important issue in the last decade, and some of the key documents include [16]:

- DoD Net-Centric Data Strategy (NCDS), May 2003
- Data Sharing in a Net-Centric Department of Defense, Dec 2004
- Guidance for Implementing Net-Centric Data Sharing, Apr 2006
- DoD Command and Control (C2) Strategic Plan Version 1.0, Dec 2008
- Interim Guidance to Implement NCDS in the C2 Portfolio, Mar 2009
- DoD C2 Implementation Plan Version 1.0, Oct 2009.

The DoD Net Centric Data Strategy (NCDS) [17] and the Army Data Transformation (ADT) [18] effort are two examples of strategy developed in this area. Both documents are designed for a larger community than C2, which is considered one Community of Interest (COI). However, both directly affect the direction of current and planned C2 systems.

The NCDS defines seven goals in its data strategy [17]:

1. Visible (who has data and what kind it is) - Data is discovered through search of catalogs, registries, etc, and visibility is accomplished through use of metadata descriptions.
2. Accessible (where and what format) - Data is posted to storage areas where it can be obtained by others. The data is accompanied by metadata descriptions and is made available to others based on access control policy.
3. Understandable (what its meaning is) - Data syntax and its semantic meaning can be uniquely interpreted.
4. Institutionalized (what and who governs it) – Data is incorporated into standard processes and practices.
5. Trusted (trustworthy, accurate and authoritative) - The validity of the data can be assessed based on its provenance, security protection, access control and integrity.
6. Interoperable - Data can be shared among different predefined or unanticipated users or systems, supported by common data models and metadata.

7. Responsive to users' needs (applicable and timely) - Methods to accommodate user perspectives via feedback are incorporated into the data practices.

The NCDS claims that the aforementioned goals do not include data quality or accuracy considerations, but that achieving the goals should result in improved data quality and accuracy. The ADT plan is aimed at processes to improve data quality as the systems are transformed to net-centric operations. The handling of data is tightly coupled with the Army Enterprise Information Architecture (EIA) that is part of the overall Army Enterprise Architecture (AEA), so that data and architecture are separated and the implementation of data services is not described. The AEA is a service oriented architecture that deals with many data oriented services such as displays (user defined dashboards), common exchange schemas, and interfaces to other C2 systems. A good description of the relationship between the Army Net-Centric Data Strategy and the Army Service Oriented Architecture is described in Reference [19]. The ADT has indicated six phases in which it is working to improve data management and data quality [18]:

1. Accountable – Incorporate common data standards and governance practices.
2. Authoritative –Identify and manage master data elements and authoritative sources.
3. Transform – Employ standardized structures and schemas such as data yellow pages to improve data sharing.
4. Expose – Make data accessible and responsive to users through the Army Data Services Layer (ADSL). Four methods of exposing data are Messaging, Data Services, Data Warehouses and Data Security.
5. Register – Validate data schemas and services against standards and then register in repositories (e.g., authoritative data repository) to enable visibility and reuse.
6. Assess – Monitor and assess data maturity levels using metrics. Measure the progress in improving data quality.

A key portion of the strategy is the ADSL, that is part of the EIA, and that provides application services for standardized handling of data, such as [20]:

- Data Mediation – Transform data among different types, vocabularies and semantics to support interoperability. Services include Structural Transform, Semantic Mediation, Data Validations and Data Brokering.

- Data Discovery and Data Access - Provide common service-based access to repositories for search and retrieval of data to support visibility and accessibility. Services include Data Search, Federated Search, Data Retrieval, Data Events and Data Streaming.

- Data Abstraction - Make data understandable through use of metadata, establish a common taxonomy and manage authoritative sources. Services include Metadata Discovery, Metadata Publishing and Data Abstraction.

- Data Management - Provide the persistence and stewardship of data resources to establish trusted data. Services include Data Replication, Data Archival, Data Auditing, and Reference Data Management.
- Data Governance - Capture and govern data resources. Services include Namespace, Schema and Ontology Management.

The ADSL hides the details of the lower layers of data handling, such as databases and repositories, from the applications and users to enable improved data portability. The connection with the data quality goals of the NCDS is clear and the use of standardized services for data handling is a strong basis for implementing data quality improvement and sustainment efforts.

In the following Table 1, we present an initial comparison mapping from the data quality concepts of ISO 8000, ISO/IEC 25012, the NCDS goals (and the ADT phases), and the intelligence community to the TMDQ sixteen dimensions.

TDQM	DoD NCDS Data Goals	Intelligence Community	ISO 8000	ISO 25012
Intrinsic:				
Free-of-error	Trusted	Accuracy	Accuracy	Accuracy, Precision
Reputation	Trusted (Accountable Authoritative)		Certification	
Believability	(Accountable Authoritative)		Certification	Credibility
Objectivity (Provenance)	Trusted (Accountable Authoritative),	Objectivity	Provenance	Traceability
Operational (Accessibility):				
Accessibility	Visible, Accessible (Expose)	Usability		Accessibility, Availability, Portability, Recoverability Performance
Security (Access Control)	Trusted (Expose)			Confidentiality
Contextual:				
Amount of Information				
Relevance	Responsive to Users Needs	Relevance, Readiness		
Value added				
Timeliness	Responsive to Users	Timeliness		Currentness

	Needs			
Completeness			Completeness	Completeness
Representational:				
Understandability	Understandable	Usability	Master Data encoding, Open Tech. Dict.	Understandability
Conciseness				
Ease of operation				Performance
Interpretability	Interoperable	Usability	Master Data Syntax	
Consistent Representations	Institutionalized, Interoperable (Transform, Register)		Master Data: Conformance	Consistency. Compliance

Table 1: Comparison of Data Quality Characteristics

In the NCDS column of Table 1 we have indicated in parentheses the phases of the ADT where they may be expected to have the most impact to data quality. For the intelligence community, it appears that usability covers several areas and would be difficult to measure objectively. Also, interestingly, the TDQM list does not seem to capture the IC notion of readiness, which indicates that the data should be adaptable to changing circumstances and requirements. The ISO 8000 and related standards provide a broad range of coverage. However, they do not address some important issues, such as timeliness or ease of operation. The NCDS fails to address certain properties, particularly timeliness, which are critical to C2. Also, although the table has indicated that NCDS covers some areas such as believability and reputation, the extent of this coverage, which is primarily limited to using authoritative data sources that have been vetted, does not span many of the situations frequently encountered in C2, such as data from a variety of sources with varying pedigree (provenance, reliability, etc). The NCDS notion of Assessment is not well captured in TDQM but is an important factor in maintaining data quality.

Other studies for various specific application contexts have identified many additional characteristics, such as a study of data quality for web portals, which identified forty-two different quality features [21]. However, we are primarily seeking to use these characteristics as an organizational tool to consider the major issues with data in C2 systems, as opposed to compiling an exhaustive listing of all possible attributes.

3 Metrics and Tools

It is useful to employ metrics to quantify the quality of the data under consideration and to make economic or strategic decisions on how to improve or maintain a given quality level. Researchers have proposed a variety of metrics that can generally be divided into objective and subjective measures, but their interpretation is typically context dependent. For instance, in some applications such as digital voice, it is acceptable to have a percentage of missing data without appreciably degrading the quality. In other applications, a missing value could be catastrophic.

In Reference [6] metrics for the sixteen TDQM features are defined as three basic forms: 1) simple ratio, 2) min or max and 3) weighted average. The metrics are typically normalized between 0 and 1. Using a simple ratio, it is possible to represent completeness, accuracy, precision, consistency, concise representation, relevancy and ease of manipulation. For example, an accuracy metric can be a simple ratio of the number of accurate records divided by total number of records. The criteria for acceptable accuracy are a function of the context or application. These metrics can be defined for high-level notions but may be made more specific to satisfy the circumstances, such as schema, column and population completeness in a database.

Min or max operations can be used for metrics that are composed of several underlying dimensions. Examples include believability, timeliness, accessibility or amount of data. For example, timeliness has been defined [22] as max [0 , 1 – (*age at delivery*/*shelf-life*)]; where

age at delivery is the delivery time minus data creation time;

shelf-life (volatility) is the total length of time that data is valid and usable;

If the age is less than the shelf life, then the data is still usable. The earlier the data is delivered, the more time there is to process the data and thus, the larger the metric. In other studies, other functional forms to represent the decay of timeliness are employed and the function is often weighted by an exponent to magnify the effects of the timeliness.

The weighted average metrics are used if there is enough detailed information on the underlying features to determine their relative contributions. In addition, weighting the simple measures can allow incorporating notions of criticality, utility and/or costs.

Some metrics are naturally objective and others subjective. “Believability,” for example, is subjective and must be assessed from user opinion or surveys rather than direct measurements or observations. In Reference [6], metrics were developed for each of the TDQM dimensions based on subjective and objective surveys of both users and system. The exact forms of the metrics or the weighting of the metrics depend on the various contextual situations. For example, timeliness may be more critical in some applications than in others. An interesting observation made in Reference [6] was that the subjective results often differed depending on the perspective of those interviewed. For example, the believability of the data was often different between the users and the data system owners. Discrepancies such as this indicate further analysis may be necessary to discover the underlying data properties.

There are many tools available in the commercial and open-source domains to support data quality measurement and improvement. Data validation tools examine data as it is input into the system and reject or correct data item errors. Extract-Transform-Load (ETL) tools can sometimes be configured to perform validation functions as the data is prepared for export and entered into an existing data set. Data profiling or data auditing tools examine a data set to identify problems such as missing, duplicate, inconsistent and otherwise anomalous data, and also to compute data quality metrics. Data cleansing (or scrubbing) tools go through an existing data set and attempt to detect, correct or remove troublesome data items (incorrect, incomplete, inaccurate, etc.). Many variations are in the market with some tools using complex reasoning and rules on relations to correct data sets. Data cleansing can be quite time consuming on large data sets and efficiency is a key consideration. Data monitoring tools are used to maintain the data quality over time as the data set is used.

It is well known that one-time attempts to improve data quality are not sufficient because data degrades over time due to factors such as data change, system change and migration. For example, data on people can change rapidly due to change of residence, death, marriage, divorce and so forth. It is generally accepted that a continual process to monitor data quality is necessary. Also necessary is clearly defined policies and governance. Several methods have been proposed to help organizations manage data quality continuously in order to achieve desired levels. One popular method, based on a diagrammatic scheme called Information Production Maps (IPMs), models data as a product that goes through manufacturing stages similar to an actual physical product and applies similar quality management procedures [23]. IPMs are particularly useful for dynamic decision environments such as an e-business, or C2 systems, where timely quality information can have a large impact on effective decision-making.

4 C2 Systems

Each of the U.S. armed services maintains its own family of C2 systems that are tailored to their particular mission needs: air, ground, sea, space, special operations. In joint and coalition operations, each participating service or nation comes with its own C2 systems. U.S. joint commands employ C2 systems that must combine information from the multiple services. Coalition commands must exchange information among the services and with C2 systems from other countries. These information-sharing requirements bring up significant problems of how to properly control access to data, and often how to control data crossing security classification domains (multilevel and cross domain security).

The functions of a C2 system are many and varied. In order to better understand where C2 fits in the warfighting domain, it is instructive to look at the U.S. Joint Staff's Joint Capabilities Areas (JCAs), a collection of the primary functions involved with warfighting [24]. C2 is one of the top-level capabilities. The nine JCAs are:

- Force Application
- Logistics
- Protection

- Force Support
- Corporate Management and Support
- Command and Control
- Battlespace Awareness
- Net-Centric
- Building Partnerships

Within Command and Control, the following capabilities are defined: *Organize*, *Understand*, *Planning*, *Decide*, *Direct* and *Monitor*. As can be inferred from these functions, C2 capabilities are heavily dependent on the quality of the information that is immediately available or that can be obtained from other sources, and also on the ability to communicate that information to and from the other capabilities. The communication functions are heavily used by the C2 functions, but are primarily included under the Net-Centric JCA and will be briefly considered further in this paper. Specific requirements for information can be issued from C2 to other JCAs such as Battlespace Awareness or Logistics, for which many of the data quality issues equally apply.

From a C2 perspective the key data issues that are frequently discussed include: interoperability, distributed access, timeliness, accuracy, provenance and security. There are also issues with information overload, as the volume of data that is available, both from the tactical and strategic sides, is rapidly increasing. The data needs to be processed in a timely manner, incorporated into the common operating picture, and delivered where needed. There are also issues associated with limited or disadvantaged communications capabilities. This limits data availability, and C2 systems must accommodate these resource-constrained situations. Looking at this from the data quality perspective, we see that most of these issues are covered by the data quality properties discussed previously. However, a data quality strategy specifically for C2 should emphasize and tailor these dimensions.

There are several cases of dramatic effects that are at least partially due to C2 data quality problems. The unintentional 1999 bombing of the Chinese Embassy in Belgrade by U.S. planes, while admittedly caused by a systemic failure in the targeting process, was plagued by data issues [25]. One example was the use of older map data that failed to show the updated location of the embassy after a move in 1996. Also, the actual address of the intended target (a warehouse) was only estimated, and not carefully verified against a map with accurate address information. Other problems were caused by duplicate target requests that appeared to come from different sources but were ultimately from the same source (this is sometimes called "ringing," and is due to a lack of provenance). Further, there was a failure to check the target against a database of known off-limits targets.

Data quality issues have also been identified in two other disasters: the space shuttle *Challenger* explosion on January 28, 1996 and the shooting down of an Iranian Airbus by the USS *Vincennes* on July 3, 1988 [26]. The Presidential Commission investigating the Challenger disaster cited flawed decision making surrounding the possible problem with O-rings at cold temperatures. The attack on the Iranian Airbus was also attributed to flawed decision making under time pressure, when the ship identified the airbus as a hostile military jet in attack mode. From the data quality perspective the decisions were affected by lapses in accuracy, completeness, consistency, relevance and fitness-for-use in the *Challenger* case, and accuracy, completeness, consistency, fitness-for-use and timeliness for the USS *Vincennes*. For the space shuttle *Challenger*, the data needed for proper analysis was available but not properly used and not presented in a form that assisted the management to make correct decisions. For the *Vincennes*, the initial misclassification occurred when users did not realize that the system reused a target designation number and then failed to resolve the resulting inconsistencies. Given all the pressures of decision making, it is arguable that data issues contributed to the erroneous decisions.

A case study of Operation *Anaconda* [27], [28] in which the US Army successfully defeated Al Qaeda forces in the Shahikot Valley of eastern Afghanistan in March 2002, showed many problems that can be partially attributed to C2 data quality. Though the operation ultimately succeeded, the initial battle plan required extensive modification. It was designed to last for a week; however, the battle lasted seventeen days, and resistance was much stronger than anticipated, requiring much more air support. Some of the problems were related to the quality of the intelligence data, such as inaccurate and incomplete estimates of enemy forces and their willingness to fight, or the disposition of civilians. There were also interoperability problems among and between joint and coalition forces. The intelligence data, which relied primarily on human intelligence, proved to be faulty and was not properly verified and vetted, reflecting believability and accuracy issues. The satellite imagery was often three days old. Some of the interoperability issues arose from a lack of unity of command, due to the relative newness of the Army forces in the area, and lack of command authority over Special Forces, air support and Afghan allied forces that were all part of the operation. For example, “Army personnel could use their FM radios to communicate directly with overhead Navy and Marine Corps aircraft but not USAF aircraft, such as F-15Es and bombers.” Also, U.S. gunships mistakenly fired on an allied Afghan column, partially causing them to turn away from the area. Although communications reportedly worked for each U.S. service component, problems occurred in communicating with other services and with allied Afghan forces. In addition, long-range communications between headquarters and edge forces was bandwidth-limited, and communication between headquarters and central command was inconsistent (timeliness, accessibility). Also, a lack of common understanding about the differing rules of engagement and procedures governing Close Air Support (CAS) contributed, reflecting understandability problems [29]**.**

4.1 Interoperability

As observed in the previous examples, the ability to share and exchange data between various C2 systems constitutes a serious problem in the C2 environment. Currently, each of the services have their own C2 systems which themselves consist of a family of related systems. In addition, the GCCS, a family of C2 systems, includes over 200 systems or services and is intended to have world-wide reach and incorporate components from all service branches [30]. Data must be exchanged among the systems in the same family as well as with other non-family systems. This problem has been well known for many years in joint and coalition settings [31] and several key developments have been achieved, such as the NATO Network Enabled Capabilities (NNEC) Common Operating Picture (COP). The NNEC COP addresses issues such as standards, dynamic tailoring, multi-level security, provenance, and knowledge management (timeliness and access). Within the U.S. government, the Universal Core (UCore) [32] is being promoted as a standard for information exchange between systems. DoD has agreed that all of the services shall use UCore (currently version 2.0) as the basis for semantic representation of data exchanges, including C2 systems data. UCore is appropriate for information exchange between the Army and other military, government agencies, non-government organizations (NGOs), and the various multinational communities (should they adopt the UCore messaging specification). UCore is an information exchange specification and implementation profile that defines a vocabulary of commonly exchanged concepts covering *who*, *what*, *when* and *where*. There is a syntactic representation based on XML, guidance for extensions for representing domain (or COI) areas, security markings, and a messaging framework. A very general taxonomy is defined to represent basic concepts, but UCore's generality needs to be tailored for each domain. Semantic layer issues for UCore, as defined in the UCore-SL (Semantic Layer) such as temporal relationships and allowing items to be of different types at different times (e.g., weapon, cargo, etc) are still being investigated by researchers.

There are other representations of the C2 domain that have been in use for some time. In particular, there is the Joint Consultation, Command and Control Information Exchange Data Model (JC3IEDM) that is in use by many countries and also by NATO. JC3IEDM exchanges are not XML-based internationally, but JC3IEDM is the Army's chosen data model for information exchange as per Reference [33].

A DoD high level data model for C2, called C2 Core, is being proposed as an extension to the C2 domain for UCore [34]. C2 Core has a C2 conceptual model and vocabulary that represents six elements of C2 systems:

1. Force Structure, Integration, Organization
2. Situational Awareness
3. Planning and Analysis
4. Decision Making and Direction
5. Operational Functions and Tasks
6. Monitoring Progress (Assessing)

There is ongoing work required to harmonize the various efforts to standardize concepts, data models and ontologies for C2. One observation is that the breakdown of C2 elements differs slightly from the JCA capabilities mentioned earlier. In a recent study of C2 data-related issues [35], it was noted that the C2 community could benefit from use of UCore and C2 Core coupled with additional C2-specific extensions to facilitate data sharing within the C2 community and the definition of core C2-specific services. The joint extensions to the C2 Conceptual model and vocabulary, the inclusion of real-world operational needs, the JC3IEDM artifacts, artifacts from ongoing data exchange development, and legacy message formats all need to be accommodated in the UCore extension. Several other key issues were identified, such as lack of a run time component and a highly complex underlying model that is not easily implemented in a modular fashion. Reference [36] presents an analysis of how to move forward with integrating the various data models. The authors conclude that UCore requires extensions to include the full JC3IEDM and that there will still need to be a more complete mapping of JC3IEDM to C2 Core. JC3IEDM is much more detailed then what is currently proposed in the C2 Core and will require the stakeholder user groups to agree on a consolidated representation that conforms to the UCore directive. Some of the implementation and runtime issues in data sharing are addressed in [16] where there is a description of the C2 Information Sharing Framework. Many of the actions and specific services designed to improve quality are described, such as adoption of UCore, C2 Core, metadata, data monitoring and data access control as well as optionally, reputation services.

Other worthwhile future interoperability developments include devising methods to enable operators to discover, use and manipulate data in ways that cannot be imagined a priori, and to do so dynamically while deployed. These capabilities are desperately needed by edge users involved in fast moving, dynamic situations. There is also a great need for data mediation services to enable system coupling and to fast-track warrior requests for data sharing. Another key scientific issue relating to interoperability involves exploring automated methods to resolve differences among the semantics of the differing systems. Even with standardized data exchange methods, there will be subtle interpretations of data that will need to be resolved. There are too many relationships among the data for people to represent and capture all of the relations between the involved entities, and the overall process would benefit if it could be automated.

4.2 Volume of Data

It is well known that the amount of raw and processed data that is entering C2 systems is growing rapidly. With the expected additions of more and more sensors, each with greater ability to produce data, the amount of raw data will explode. Even now, in some surveillance applications, data is being generated at a faster rate than can be processed, and it ends up being archived for later examination. With the increasing use of unmanned platforms, such as Unmanned Aerial Vehicles (UAVs), the demand for information delivered in real time to the edge is also growing. Consider the Air Force Reaper-mounted “Gorgon Stare” which can transmit up to 65 video images per second [37], or future systems such as the Defense Advanced Research Projects Agency (DARPA) Autonomous Real-time Ground Ubiquitous Surveillance System (Argus IS) platform [38] with 1.8 Giga-pixel video sensors generating data at 27 Gigabits per second. Such systems can quickly overwhelm the ability of C2 systems to process the information. As a result of this increase, many intelligence, surveillance and reconnaissance (ISR) decision support systems are receiving large volumes of data with poor control of data quality (e.g., noise, clutter). Requests requiring adaptable analysis methods and unpredictable data requirements are normal occurrences. For example, tracking vehicles in an urban environment or identifying placement of roadside bombs from video are typical examples of particularly challenging requests. There is also the problem of short and long term storage, data accessibility and supplying the computational power to process the requests. In this context, timeliness becomes a critical property. If the raw or processed data is not available to track a target, then it quickly becomes of limited value.

There are a wide variety of scientific and technical challenges relating to handling large volumes of data. These include novel architectures for storing and accessing large data sets, processing architectures to analyze the data, and methods for securely sharing data and results. Management of large data sets, including multi-level classifications, is a challenge. Various research programs are being formulated, to address many of the scientific challenges in these types of issues. At least twenty-six research projects related to commander's decision support systems have been identified in 2009 [39]. Other newer projects, such as the Data-to-Decisions project, are focused on handling the volume of data issue. Some analysts have suggested that greater emphasis should be put on assisting users to understand information rather than designing for full automation. However, in either case, additional emphasis should be given to understanding the effects of large volumes of data on the data quality dimensions, such as availability and usability and incorporating this into the decision processes.

4.3 Trustworthiness of data

For C2 systems, determining the level of trust to place in data can be extremely important. It is often difficult to determine whether separate reports are referring to the same or different incidents, Detecting data ringing, where the same report is relayed by different individuals, can be a serious challenge. Similarly, copy-paste is frequently used in report generation and automated tracking of sources from copy-paste operations would be very helpful in determining trust. Incorporating some form of provenance data is needed to help clarify these situations. Within the DoD, the services are currently focused on defining Authoritative Data Sources (ADSs) and using a standardized metadata registry for data discovery and use. These systems have limited provenance data, primarily containing only the source and date. Outside of the authoritative sources, there is almost no provenance tracking.

In the emerging research, provenance data should contain all the information necessary to determine the complete history of the data. For certain applications, such as bioinformatics or physics, it is appropriate to capture the entire workflow that transformed the data from input to output for purposes of validation or repeatability [40-41]. Other applications mainly require documentation of original sources, context or other relevant pieces of information. In [42] a W7 model (*What*, *Who*, *When*, *Where*, *Which*, *How*, and *Why*) is given that captures all relevant information for full documentation of a data life-cycle from creation to destruction, however, this can require huge amounts of storage in practical scenarios. There are many research activities working towards automating the capture of provenance data with techniques for special situations such as copy-paste, database, grid computing systems, file systems, service oriented architectures, enterprise service bus and archiving systems. For resource-limited environments, such as those often faced by C2 systems, there are limits to the amount of provenance that can be collected, stored or transmitted. Further research is needed to characterize the utility of provenance models for the various C2 scenarios.

5 Conclusions

We have described the characteristics of data quality from several different communities such as academia, the commercial world, the IC, and the DoD, and described a set of data quality attributes, primarily based on the TDQM sixteen dimensions of data quality. The IC's "usability" criteria cover several different concepts that are subjective and difficult to measure, while the DoD's NCDS arguably does not adequately address the notion of data timeliness. The NCDS covers some important factors such as believability and reputation, but the coverage is primarily limited to using authoritative, vetted data sources. This does not address important situations where data comes from a variety of sources with varying degrees of reliability. On the other hand, the TDQM criteria do not adequately capture the notions of readiness and adaptability.

C2 systems are beset with similar data quality issues as is the general enterprise IT community and all of the data quality characteristics are relevant to C2 systems. However, several quality issues are of relatively greater importance to C2 systems because of the potential lethality of decision-making errors. These characteristics are not independent and they should not be addressed in isolation, but should be part of an ongoing data quality enhancing process. Research and development is needed to determine how best to accomplish this within the constraints of real-time decision making environments with limited bandwidth, processing power and intermittent service in a disruption-tolerant and robust fashion. The resulting benefits include improved decision making through explicit use of the data quality features, such as believability, provenance or reputation.

There are several specific steps that need to be taken to improve data quality in C2 systems:

- Examine several specific C2 systems in greater detail for data quality characterization and requirements, in order to discover the tradeoffs in C2 contexts.
- Define and develop a set of C2-specifc data quality dimensions, metrics and associated weightings. Particular emphasis should be on accuracy, timeliness and trust. This would have to be part of a trade-off analysis when there are limited resources, based on the benefit provided by the data-quality information.
- Incorporate current standards, such as ISO 8000 and 25012, tailored for C2 applications, into C2 systems.

- Set policies and governance to accomplish data quality goals and methods of enforcement.
- Define standard data and metadata services to provide a consistent environment for data handling. The ADSL is an excellent start at this process.
- Data quality characteristics and their associated metrics should be explicitly incorporated into C2 systems and their operations with ongoing procedures for monitoring and governance enforcement.
- Incorporation of some forms of provenance information coupled with the data items, beyond designating certain repositories as authoritative data sites, is needed to improve the credibility of decisions. Decision support systems should be enhanced to incorporate this information.
- Interoperability and data sharing should continue to be addressed by the C2 community at the syntactic and semantic levels, and these standards should be implemented across all current and planned C2 systems. More rapid methods of modifying and updating the data exchange standards needs to be developed. In addition, some of the tenets, such as "publish first," may need to be rethought and revised in terms of data quality impact.

Even with the implementation of the above steps, there remain many unresolved issues and areas of active research on general aspects of data quality. There is active research on automated provenance handling, but it remains a challenging problem. It remains very difficult to determine if a document or web page has been copied or combined from other sources, unless it has been under version control for its entire existence.

Few results have been reported on research specifically directed to C2 systems. There are various types of C2 data to be considered when capturing C2 data quality features and the quality requirements of raw data may be very different from a command message or a situation report. One approach to incorporating data quality capabilities is to provide appropriate metadata along with every data item so that the data becomes self-describing and self-protecting. Methods to best accomplish this within the constraints of resource-limited C2 systems still need to be explored.

Current practices in C2 involve the human-in-the-loop for almost all levels of data entry and analysis. The increase in data volume is overwhelming both the people and the systems and causing mistakes induced by information overload. Increased use of machine processing of the raw data and elementary decision making is necessary if modern commanders are to operate effectively under this data deluge. The commanders must be involved at the crucial decision points and provided with situation awareness, but otherwise not encumbered by the lower level data details. Decision support systems that can process raw data and make the low-level decisions, alerting the commanders at the crucial points, should be investigated. Incorporation of data quality characteristics along with other forms of metadata that are semantically defined and can be processed and understood by the decision software may go far in facilitating this environment. Recent research in semantic characterizations that incorporate metadata as additional data in the knowledge base, so that the metadata can be accessed, manipulated and used in further inferences, are an alternative to more traditionally structured relational data repositories. These environments can naturally incorporate quality features and use them to assist the decision maker in understanding the credibility of the information relied upon.

Examining C2 systems from the data-quality point of view provides an alternative and comprehensive cross section of relevant system performance attributes. By enabling C2 systems to explicitly (and in the future, automatically) take the quality dimensions of the underlying data into account, the decision-making ability of the commanders will be enhanced by reducing the uncertainty in their decision space.

References

1. Sebastian, M., *The Cost of Dirty Data to Accounts Receivable Managers*, in *Inside Accounts Receivable Management*. 2008.
2. ISO/IEC, *ISO/IEC 2382-1:1993 Information technology -- Vocabulary -- Part 1: Fundamental terms*. 1993, ISO/IEC.
3. Simon, A.J., *Overview of the Department of Defense Net-Centric Data Strategy*. Crosstalk The Journal of Defense Software Engineering, 2006
4. Strong, D.M., Y.W. Lee, and R.Y. Wang, *Data Quality in Context*, in *Communications of the ACM*. 1997. p. 103-110.
5. Eppler, M. (2010) *IAIDQ Glossary*.
6. Pipino, L.L., Y.W. Lee, and R.Y. Wang, *Data Quality Assessment*, in *Communications of the ACM*. 2002, Association of Computing Machinery. p. 211-218.
7. ISO Technical Committee TC 184, A.s.a.i., sub-committee SC 4, Industrial data, *ISO/TS 8000-100:2009 Data Quality - Part 100: Master Data (Overview)*. 2009.
8. ISO Technical Committee TC 184, A.s.a.i., sub-committee SC 4, Industrial data, *ISO 8000-102:2009 Data quality — Part 102: Master data: Exchange of characteristic data: Vocabulary*. 2009, ISO/IEC.
9. ISO Technical Committee TC 184, A.s.a.i., sub-committee SC 4, Industrial data, *ISO/IEC 8000-110:2008 Data Quality - Part 110: Master Data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification*. 2008, ISO/IEC.
10. Grantner, E., *8000 - A Standard for Data Quality*, in *Logistics Spectrum*. 2007.
11. ISO Technical Committee TC 184, A.s.a.i., sub-committee SC 4, Industrial data, *ISO 22745-1:2010 Industrial Automation Systems and Integration - Open Technical Dictionaries and their application to master data - part 1: Overview and fundamental principals*. 2010, ISO/IEC.
12. Association, E.C.C.M. *Electronic Open Technical Dictionary (eOTD).* 2010; Available from: http://www.eccma.org/index.php.
13. ISO/IEC Technical Committee JTC-1/SC 7, S.a.S.E., *ISO/IEC 25012:2008 Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model*. 2008, ISO/IEC.
14. Brei, W.S., *Getting Intelligence Right: The Power of Logical Procedure. Occasional Paper #2*. 1996, Joint Military Intelligence College (JMIC).

15. Zhu, H. and R.Y. Wang, *Information Quality Framework for Verifiable Intelligence Products*, in *Data Engineering: Mining, Information, and Intelligence*, J.R.T.a.T.T. Y. Chan, Editor. 2007, Springer.
16. Walsh, P. and et al., *DoD Net-Centric Services Strategy Implementation in the C2 Domain*. 2009, Institute for Defense Analyses.
17. Officer, D.C.I., *Net-Centric Data Strategy*. 2003, U.S. Department of Defense.
18. Army, U.S. *Army Net Centered Data Strategy*. 2010 [cited 2010 Dec. 2010]; Available from: http://data.army.mil/.
19. Dirner, M., E. Yuan, and J. Blalock, *Realizing the Army Net-Centric Data Strategy (ANCDS) in a Service Oriented Architecture (SOA)*. 2008, George Mason University.
20. Army, U.S. *Army Data services Layer (ADSL)*. 2010 [cited 2010 Dec. 2010]; Available from: http://data.army.mil/ADSL.html.
21. Moraga, C., et al., *SQuaRE-Aligned Data Quality Model for Web Portals*, in *2009 Ninth International Conference on Quality Software*. 2009, IEEE Computer Society. p. 117-121.
22. Ballou, D.P., et al., *Modeling Information Manufacturing Systems to Determine Information Product Quality*. Management Science, 1998. **31**(2): p. 462-484.
23. Shankaranarayan, G., M. Ziad, and R. Wang, *Managing Data Quality in Dynamic Decision Environments: An information Product Approach*. Journal of Database Management, 2003: p. 14-32.
24. J-7/JFDID, J.S. *Joint Capability Area Management System (JCAMS)*. 2010; Available from: http://jcams.penbaymedia.com/.
25. Myers, S.L., *Chinese Embassy Bombing: A Wide Net of Blame*, in *New York Times*. 2000: New York, NY.
26. Fisher, C. and B. Kingma, *Criticality of Data Quality as Exemplified in Two Disasters,*. Information and Management, Elsevier Science, 2001. **39**: p. 109-116.
27. Kugler, R., *Operation Anaconda in Afghanistan: A Case Study of Adaptation in Battle*, in *Case Studies in Defense Transformation*. 2007, National Defense University, Center for Technology and National Security Policy: Fort Lesley J. McNair, Washington, DC, 20319.
28. MacPherson, M., *Roberts Ridge : A Story of Courage and Sacrifice on Takur Ghar Mountain, Afghanistan*. 2005: Dell Publishing.
29. Kugler, R., M. Baranick, and H. Binnendijk, *Operation Anaconda: Lessons for Joint Operations*. 2009, National Defense University, Center for Technology and National Security Policy: Fort Lesley J. McNair, Washington, DC, 20319.
30. Ceruti, M., *Data Management Challenges and Development for Military Information Systems*. IEEE Transactions on Knowledge and Data Engineering, 2003. **15**(5): p. 1059-1068.
31. Waters, J., B. Powers, and M. Ceruti, *Global Interoperability Using Semantics, Standards, Science and Technology (GIS3T)*. Computer Standards and Interfaces, Elsevier, 2009. **31**: p. 1158-1166.
32. Government, U.S. *Universal Core*. 2010; Available from: https://ucore.gov/ucore/.
33. Staff, O.o.t.D.C.o., *DAMO SBB Memo: Command and Control Information Exchange Data Model*. 2005, U.S. Department of the Army.
34. Allen, M.D., C. Macheret, and M.A. Malloy, *C2 Core and UCore Message Design Capstone: Interoperable Message Structure*. 2009.
35. Guthrie, P. and et al., *Independent Assessment Team Report on C2 Data*. 2008, Institute for Defense Analyses.
36. Haugh, B., et al., *JC3IEDM, C2Core, and UCore: A White Paper*. 2009, Institute for Defense Analyses.
37. Kenyon, H., *New Surveillance systems offers wide-angle view of battlefield*, in *Defense Systems*. 2011.
38. Robinson, B., *New UAV sensors could leave enemy no place to hide*, in *Defense Systems*. 2009.
39. Lau, C., *Data-to-decision notes*. 2010.

40. Cohen, S., S. Cohen-Boulakia, and S. Davidson, *Towards a Model of Provenance and User Views in Scientific workflows.* Lecture Notes in Computer Science, 2006. **4076**: p. 264-279.
41. Moreau, L., et al., *The Provenance of Electronic Data.* Communications of the ACM, 2008. **51**(4): p. 52-58.
42. Ram, S. and J. Liu, *A Semiotics Framework for Analyzing Data Provenance Research.* Journal of Computing Science and Engineering, 2008. **2**(3): p. 221-248.