An Information Quality Framework for Verifiable Intelligence Products

Hongwei Zhu[1], Richard Y. Wang[2]

[1]College of Business and Public Administration, Old Dominion University
[2]MIT Information Quality Program, Massachusetts Institute of Technology

The intelligence community (IC) gathers and analyzes a large amount of information from a variety of sources to produce intelligence products for decision makers and policy makers. It is imperative that these intelligence products are of highest quality. From raw inputs to final intelligence products, there are often many processes involved. In addition to the quality of final intelligence products, another important aspect of quality is the *verifiability* of the final intelligence products: how is the conclusion derived, from what sources? In this chapter we will illustrate the need for verifiability and propose a solution towards Variable Intelligence Products (VIP). We will use terms *information quality* (IQ) and *data quality* interchangeably in the rest of the chapter.

1 Introduction

Organizations have been increasingly investing in technology to collect and process vast volumes of data. Even so, they often find themselves stymied in their efforts to effectively use the data to improve business processes and to make better decisions. This difficulty is often caused by information quality issues within the organization and other related organizations.

The government is not immune to the problems of IQ. In the U.S., poor data quality has caused serious problems. In response, many government agencies have begun to establish IQ policies and guidelines as part of their effort in practicing Enterprise Architecture (EA) to improve collaboration and information sharing across the Community.

This chapter presents an effort to improve IQ within the broader IC EA initiative. Its motivation can be illustrated by a hypothetical example (Fig. 1) adapted from "A Compendium of Analytic Tradecraft Notes" issued by the Directorate of

Intelligence of the Central Intelligence Agency (CIA Directorate of Intelligence, 1997).

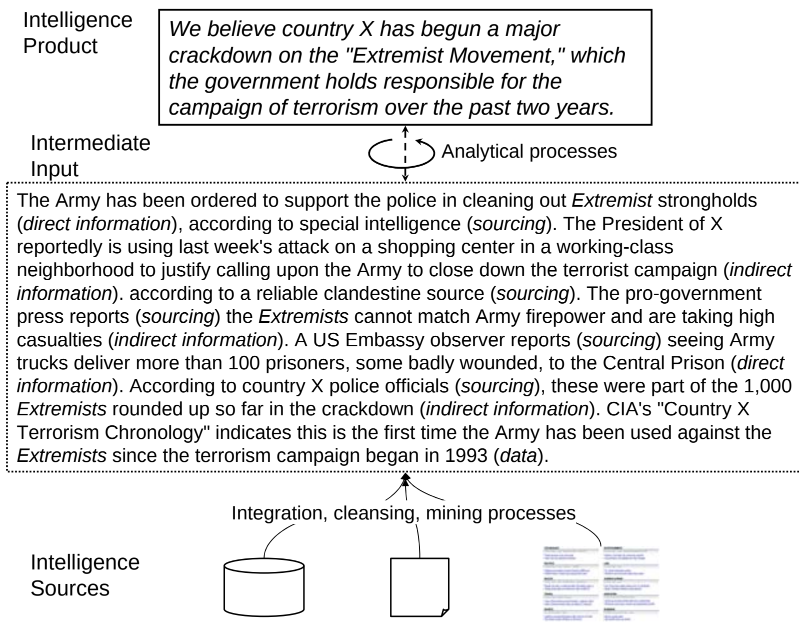

Fig. 1. From Intelligence Information to Intelligence Product

A statement in an intelligence product is shown in the top of Fig. 1. It is derived from the intermediate input shown in the middle of Fig. 1 via certain analytical processes. The intermediate input is obtained from various sources, shown in the bottom of Fig 1, using such techniques as information integration/extraction, information retrieval, data cleansing, and data mining. Two questions the customers of the intelligence product may ask are:

- How good is the statement (i.e., what is the overall quality of the product)?
- How is the statement derived (i.e., what are the sources and procedures that lead to the statement)?

The first question concerns with quality assessment of the intelligence product. The second question concerns with the capability of verifying sources and explaining processes of intelligence production. We will develop an information quality framework to address both questions. The framework integrates key findings from two decades of information quality research and adapts them to take into account of the characteristics of the IC. Specifically, we will develop a comprehensive and systematic set of metrics for measuring the quality of intelligence products. We will also develop techniques for tracing the sources and production processes involved in producing the final intelligence products.

The rest of the chapter is organized as follows. Section 2 provides background information about the production of intelligence products and key findings IQ research. Section 3 discusses IQ challenges faced by the IC. Section 4 presents a proposed solution. Section 5 concludes the chapter.

2 Background

2.1 Production Process of Intelligence Products

Intelligence production is a dynamic and iterative process, as illustrated in Fig. 2. Below we describe the major steps involved.

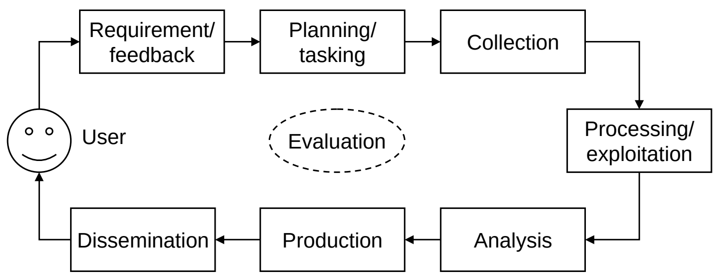

Fig. 2. Intelligence Production Process (adapted from Krizan, 1999)

Requirements

Intelligence customers often express their needs using their own terminology. Such needs are converted to intelligence requirements understood within the IC. This is often done by asking the five-W questions: Who, What, When, Where, and Why. Sometimes the intelligence requirements can be characterized using a taxonomy of problems (Jones, 1995) that classifies problems into five categories: simplistic, deterministic, moderately random, severely random, and indeterminate. The expected quality of products varies amongst products that deal with different types of intelligence problems. A method for describing IQ is needed.

Planning and Collection

Sometimes the intelligence at hand is sufficient to meet the customer requirements and no extra intelligence needs to be collected. This requires the ability of knowing what information is available. Search technology and information retrieval tools can be helpful. At other times, intelligence collection is necessary; in this case the collection phase of the intelligence process involves several steps:

translation of the intelligence need into a collection requirement, definition of a collection strategy, selection of collection sources, and information collection.

There are four types of sources that the IC primarily uses. Their characteristics and intended uses are summarized in Table 1 (Clauser and Weir, 1975).

Table 1. Characteristics of Intelligence Information Sources

Source	Collection Disciplines and Source Attributes	Analytic Use
People	HUMINT; subject-matter experts, professional researchers, information specialists, eyewitnesses or participants	Transfer of first-hand knowledge, referral to other sources
Objects	IMINT; physical characteristics of equipment, materials, or products, such as texture, shape, size, and distinctive markings	Basis for emotive but objective reporting on composition, condition, origin, or human purpose
Emanations	MASINT, SIGINT; detectable phenomena given off by natural or manmade objects; electromagnetic energy, heat, sound, footprints, fingerprints, and chemical and material residues	Scientific and technical analysis
Records	IMINT, SIGINT; symbolic (written and oral reports, numerical tabulations) or non-symbolic (images, electro-magnetic recordings of data)	Research, background information, translation, conversion to usable form

Recently, information from open sources (OSINT) (e.g., publicly accessible websites) has become increasingly important to intelligence. An Open Source Information System (OSIS) has been developed to help the IC to gather and process intelligence information.

The quality and reliability of these different sources need to be assessed routinely to determine the quality of the finished intelligence products produced using these sources.

Processing

This is the process that transforms raw data to intelligence information. Processing methods vary depending on the form of the collected information and its intended use. One important procedure is information collation, which organizes the information into a usable form, adding meaning where it was not evitable in the original. Collation includes gathering, arranging, and annotating related information; drawing tentative conclusions about the relationship of "facts" to each other and their significance; evaluating the accuracy and reliability of each item; grouping items into logical categories; critically examining the information source; and assessing the meaning and usefulness of the content for further analysis. Collation reveals information gaps, guides further collection and analysis, and provides a framework for selecting and organizing additional information (Mathams, 1995). It is important that no bias is introduced in the selection and interpretation of information during collation.

Analysis

Three levels of analysis are often conducted for different customer requirements (Krizan, 1999):

- *Describe*: fully describe the phenomenon under study, accounting for as many relevant variables as possible.
- *Explain*: thoroughly explain the phenomenon through interpreting the significance and effects of its elements on the whole.
- *Estimate*: provide synthesis and effective persuasion about the situation.

These different levels of analysis are better known as "intelligence food chain" (Davis, 1995) within the IC:

- *Facts*: verified information related to an intelligence issue (for example: events, measured characteristics).
- *Findings*: expert knowledge based on organized information that indicates, for example, what is increasing, decreasing, changing, taking on a pattern.
- *Forecasts*: judgments based on facts and findings and defended by sound and clear argumentation.
- *Fortunetelling*: inadequately explained and defended judgments.

The mnemonic "Four Fs Minus One" may serve as a reminder of how to apply this criterion. Whenever the intelligence information allows, and the customer's validated needs demand it, the intelligence analyst will extend the thought process as far along the Food Chain as possible, to the third "F" but not beyond to the fourth (Krizan, 1999).

Commonly used methods for analysis include opportunity analysis, linchpin analysis (Davis, 1995), and analogy. It is important that no misperceptions and bias are introduced during analysis.

Production

This is the process of creating the finished intelligence products in any medium usable by the customers. Product content and tone often need to be adjusted according to the level of expertise of the customer. Caution needs to be exercised to avoid misinterpretation by intended customers as well as incidental customers.

2.2 Current IQ Practices in the IC

Within the Central Intelligence Agency, and the IC in general, the Kent doctrine has been widely adopted for quality assurance in intelligence analysis. As enumerated by Frans Bax, founding Dean of the Kent School, and discussed in Davis (2002), the doctrine consists of nine aspects:

- Focus on Policymaker Concerns
- Avoidance of a Personal Policy Agenda
- Intellectual Rigor

- Conscious Effort to Avoid Analytic Biases
- Willingness to Consider Other Judgments
- Systematic Use of Outside Experts
- Collective Responsibility for Judgment
- Effective communication of policy-support information and judgments
- Candid Admission of Mistakes

In addition to the doctrine, the IC relies on customer feedback to evaluate the quality of intelligences products. The evaluation framework presented in Brei (1996) uses six criteria to evaluate the products of the sub-processes of the production process:

- *Accuracy*: All sources and data must be evaluated for the possibility of technical error, misperception, and hostile efforts to mislead.
- *Objectivity*: All judgments must be evaluated for the possibility of deliberate distortions and manipulations due to self-interest.
- *Usability*: All intelligence communications must be in a form that facilitates ready comprehension and immediate application. Intelligence products must be compatible with a customer’s capabilities for receiving, manipulating, protecting, and storing the product.
- *Relevance*: Information must be selected and organized for its applicability to a customer’s requirements, with potential consequences and significance of the information made explicit to the customer’s circumstances.
- *Readiness*: Intelligence systems must be responsive to the existing and contingent intelligence requirements of customers at all levels of command.
- *Timeliness*: Intelligence must be delivered while the content is still actionable under the customer’s circumstances.

This framework suggests a 2-dimentional evaluation matrix as shown in Table 2 to be used to assess the quality of interim and finished intelligence products.

Table 2. Evaluation Matrix for Intelligence Product

	Needs Definition	Collection	Processing	Analysis	Production
Accuracy					
Objectivity					
Usability					
Relevance					
Readiness					
Timeliness					

The Kent doctrine and the evaluation matrix, when systematically applied, should provide means to ensure that the intelligence products are of high quality. However, there are still areas where we can improve.

2.3 Relevant Concepts and Methods of IQ Management

The Total Data Quality Management (TDQM) framework (Madnick and Wang, 1992) has been widely adopted and proven to be effective in practice. The framework consists of a set of concepts and methods for describing, measuring, and improving information quality. Our preliminary investigation indicates that the following concepts and methods are relevant to the IC:

- Multiple dimensions of IQ
- Treating information as a product
- IP-Map – a methodology for describing and optimizing information manufacturing process
- PolyGen – a model for maintaining data lineage and useful IQ attributes
- QER – a conceptual modeling technique to incorporate quality in conceptual data model

2.3.1 TDQM Framework

The TDQM framework advocates continuous data quality improvement through cycles of *Define*, *Measure*, *Analyze*, and *Improve* (Madnick and Wang 1992). The framework extends the Total Quality Management (TQM) framework for quality improvement in manufacturing domain (Deming 1982; Juran and Godfrey 1999) to the domain of data.

Define. It has been found effective to define data quality from consumer's point of view as *fitness for use* (Strong et al., 1997). Further research identified the dimensions of data quality (Wang and Strong, 1996). These dimensions are organized in four categories, as shown in Table 3. Intrinsic data quality denotes the quality that data inherently has. Accessibility and representational data quality emphasizes the role of systems that store, process, and deliver data to the consumers. Contextual data quality highlights that data quality must be considered within the context of the task at hand (e.g., imagery data with a certain resolution may be sufficient for one task but insufficient for another).

Table 3. Data quality categories and dimensions

Category	Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-added, Timeliness, Completeness, Amount of data
Representational DQ	Interoperability, Ease of understanding, Concise representation, Consistent representation

Measure. A comprehensive data quality assessment instrument has been developed for use in research as well as in practice to measure data quality in organizations (Lee et al. 2002). The instrument operationalizes each dimension into four to five measurable items; appropriate functional forms are applied to these items to determine the score of each dimension (Pipino et al. 2002). The instrument can be adapted to accommodate specific organizational needs.

Analyze. The measurement results are interpreted at this step. The analysis determines the dimensions that need improvement and the root causes of data quality problems. Gap analysis techniques (Lee et al. 2002) can be used to reveal the data quality perception gaps between different dimensions and between different roles of data production process. The three major roles are data collectors, data custodians, and data consumers (Lee and Strong 2004).

Improve. At this step, actions are taken to change data values directly or, more appropriately, change the processes that produce the data. The latter approach is more effective as discussed in (Wang 1998; Wang et al. 1998), where steps towards managing information as a product are provided. In addition, technologies mentioned earlier such as polygen and Quality-ER can be applied as part of the continuous improvement process.

2.3.2 Treating information as Product and IP-Map

To effectively improve IQ, an organization should treat information as a product instead of a byproduct. An information product (IP) needs to conform to specifications and to meet consumer expectations.

To operationalize the notion of "information as a product", we need to model the information manufacturing process. Many modeling methods for information manufacturing systems have been developed. Almost all of these lack the ability to systematically represent the manufacturing processes. The proposed information product map (IP-MAP) method can systematically model the manufacture of an IP (Shankaranarayan et al., 2003; Shankaranarayan and Wang, 2007). The IP-MAP is an extension of the Information Manufacturing System (IMS) proposed earlier. This representation offers several advantages:

- It allows the IP manager to visualize the most important phases in the manufacture of an IP and identify the critical phases that affect its quality.
- Using this representation, IP managers will be able to pinpoint bottlenecks in the information manufacturing system and estimate the time to deliver the IP.
- Based on the principles of continuous improvement for the processes involved, the IP-MAP representation would not only help identify ownership of the processes at each of these phases but would also help in implementing quality-at-source.
- The representation would permit IP managers to understand the organizational (business units) as well as information system boundaries spanned by the different processes / stages in the IP-MAP.
- It permits the measurement of the quality of the IP at the different stages in the manufacturing process using appropriate quality dimensions.

2.3.3 PolyGen

Intelligence analysis is often accomplished using information contained in heterogeneous/distributed databases. In this environment, data consumers often need to know not only the sources of information, but also the intermediate sources that helped in composing the information. A PolyGen model (Wang and Madnick, 1989) has been developed to address this issue. The PolyGen model is named for its multiple (poly) source (gen) perspective. It uses a data source tagging mechanism to answer the questions “where did the data come from” and “which intermediate sources were used to derive the data.” The model has both a data structure and a query answering mechanism to help objectively determine the quality of the data. Follow on research has developed methods for managing data lineage of semi-structured data such as XML (Buneman et al. 2001) and implemented data lineage management as a part of the query processing engine (Widom, 2005).

2.3.4 QER

QER is an extension to Entity-Relationship model to capture data quality requirements in the design phase. As illustrated in (Wang et al. 1993), this extension can capture data quality requirement as meta-data at the cell level. Furthermore, the querying system can be extended to allow for efficient process of data quality meta-data (Wang et al. 1995). A recent extension to QER can be found in Jiang et al. (2007). The QER method is important to the IC and can be incorporated into the IC-wide Enterprise Architecture efforts.

3 IQ Challenges within the IC

We have presented a motivating example earlier with two primary questions related to a range of IQ challenges faced by the IC. In this section, we review these challenges to provide a better understanding of the issues addressed by the research.

3.1 IQ Issues in Intelligence Collection and Analysis

Intelligence collection and analysis is an inexact science at best. Accurate results often require data from multiple sources or data that has been collected over time and then integrated to develop a final product. Critical elements of data may reside in the databases of multiple IC organizations. There are major problems with intelligence that must be addressed. A few of these include:

- Incompleteness. IC organizations usally cannot collect all nencessary information because of the obstacles created by the adversaires. Also, it is often difficult to validate the collected information. To address this challenge, the IC attempts to collect from multiple sources to corroborate the facts.

- Inconsistency. Information from multiple sources sometimes points to different answers. It is difficult to determine which information is correct and which is false.
- Uncertainty and validity. The adversaries often execute deception operations. It can be difficult to determine the if collected information is the true fact or deception.
- Source identification. A problem that has been an issue in human intelligenc (HUMINT) collection is the same source providing the same information to multiple collection organizations. When integrated, it appears that we have multiple sources that corroborate the facts. Worse, if the source is not trustworthy, he/she may pass invalid information to each organization. Also, the source may be passing information solely for the purpose of advancing a personal objective that may or may not be an objective of the IC.

These problems directly affect the quality of finished intelligence products. To a certain extent, these problems are related to the ability (or lack thereof) of identifying the actual sources of information and the quality of information. For example, in the case of inconsistent information, we can give more weight to information from sources of higher reliability. If we know the true source of information, we can avoid the problem of mistaking a single source as multiple sources in the case of HUMINT collection. In addition, we also need to consider the process of intelligence production since error can be introduced at any point of the process.

3.2 Other IQ Problems

In addition to the issues in the areas of intelligence collection and analysis, there are several other IQ problems within the IC. Some of the areas of concern are:

- The problem of maintaining the quality of information for which there is a known authoritative source. One example is personnel data for members of the IC organizations. The quality of this information must be maintained as it is entered into an IC directory to support information sharing among IC organizations. Faulty or duplicate data can result in denial of access to information critical for individuals to do their jobs or improper access that can compromise the data.
- The problem of maintaining the quality of information developed from the analysis of multiple sources of intelligence. Once a fact has been determined it must be recognized and maintained consistently by the members of the IC. Fairly simple areas like ensuring embassy locations are properly maintained can help eliminate political embarrassments and unwanted loss of life as happened in Serbia.
- The problem of maintaining consitency and timeliness of information as exisiting information is revised and updated. At times what is known (or thought) to be true is discovered to need modification. An IC organization may recall the information or publish modifications. These must be accurately

disseminated to the other IC organizations and used to modify their internal databases.

As the IC expands its information sharing, information exchanges and collaborative analysis, IQ areas like those above must be addressed.

3.3 IQ Dimensions Related to the IC

The various issues discussed above can be understood from the 16 dimensions of IQ. The dimensions are general enough to be used across sectors. However, when they are applied to the IC, specific considerations need to be taken into account because of certain unique characteristics of the IC. Table 4 summarize the IQ dimension definitions (Wang and Strong, 1996; Kahn et al., 2002) and the specific considerations of the IC (Mosier, 2005).

Table 4. IQ Dimensions and IC Considerations

Dimension	Definition	IC Considerations
Accessibility	The extent to which data is available, or easily and quickly retrievable.	This has been shown to be a major deficiency across the IC. Not explicitly identified by the definition, but import to the IC, is knowledge of existence of the information.
Amount of Information	The extent to which the volume of data is appropriate for the task at hand.	Difficult to guarantee. Information can be overloaded in certain cases and extremely insufficient in other cases.
Believability	The extent to which data is regarded as true and credible.	This is an extremely important dimension for the area of collection and analysis. It addresses the deception and false information areas.
Reputation	The extent to which information is highly regarded in terms of its source or content.	Information gathered from imagery usually has a high reputation, especially when multi-spectral analysis is done. Information from SIGINT is considered accurate, but may always be a deception. HUMINT is sometime suspected by the consumers.
Completeness	The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.	Challenging because of difficulties in collection. The potential for IQ processes to determine the level of completeness could be a major area of investigation.
Concise Representation	The extent to which data is compactly represented	
Consistent Representation	The extent to which the data is presented in the same format.	Major efforts in this area through IC standardization efforts.
Ease of Operations	The extent to which data is	

	easy to operate on and apply to different tasks.	
Free-of-Error	The extent to which data is correct and reliable.	Very important, but hard to ensure in the collection area because of deception and other factors.
Interpretability	The extent to which data is in appropriate languages, symbols, and units and the definitions are clear.	Increasingly important as IC organizations begin to share information.
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial.	An important area for analysis within the IC. Intelligence products that contain the analyst bias can lead to potential disaster.
Relevancy	The extent to which data is applicable and helpful for the task at hand.	
Security	The extent to which data access to data is restricted appropriately to maintain its security.	Of major importance to the IC.
Timeliness	The extent to which data is sufficiently up-to-date for the task at hand.	In the IC it is critical to get the time sensitive information to the people who need it in time.
Understandability	The extent to which data is easily comprehended.	
Value Added	The extent to which data is beneficial and provides advantages from its use.	This is an area where further investigation could lead to insights into how data integrated from multiple sources add value to a final product.

4 A Proposed Solution

In this section we propose a solution to address the questions raised in the motivating example: (1) determining the quality of intelligence products; and (2) being able to verify the sources and processes that lead to the final intelligence product. The two issues are related: the quality of the final product is determined by the quality of all original and intermediate sources. At each stage of the production process, the quality of the products needs to be evaluated. Conversely, as we trace the production process of an intelligence product, we want to examine the quality of intermediate products in the process. Thus the proposed solution consists of two components: (1) a set of metrics for assessing the quality of information products throughout the production process; and (2) a mechanism for enabling the verifiability of intelligence products.

4.1 IQ Metrics for Intelligence Products

The IC produces a variety of intelligence products for different purposes. To assure the quality of intelligence product and to continuously improve product quality, it is necessary to establish IQ metrics for these products. The metrics will provide a means of monitoring product quality over time and optimizing resources for quality improvement. At the minimum, the metrics will consist of the six attributes that have been used within the IC: accuracy, objectivity, usability, relevant, readiness, and timeliness (Brei, 1996). The metrics can be further extended to include the 16 IQ dimensions (Wang and Strong, 1996).

Establishing systematic metrics for each individual intelligence product is costly and inefficient. The inefficiency can be avoided if we can appropriately categorize the products, in which case we establish a set of metrics for each category of products. Garst (1989) suggests categorizing intelligence products according to the *subject* and the intended *use* (See Table 5**).**

Table 5. Categories of Intelligence Products

By Subject	**By Use**
Biographic	Research
Economic	Current
Geographic	Estimative
Military	Operational
Political	Scientific and Technical
Sociological	Warning
Scientific and Technical	
Transportation and Communications	

A product covering a certain subject is often used for different purposes or uses. For example, a *military* analysis can be used for *operational* purposes and for *research* purposes. The quality requirements for different purposes are often different. Therefore, a set of quality metrics should be established for each use of a particular category of products. We envisage a quality metrics system in the form a 3-dimentional matrix, as illustrated in Fig. 4.

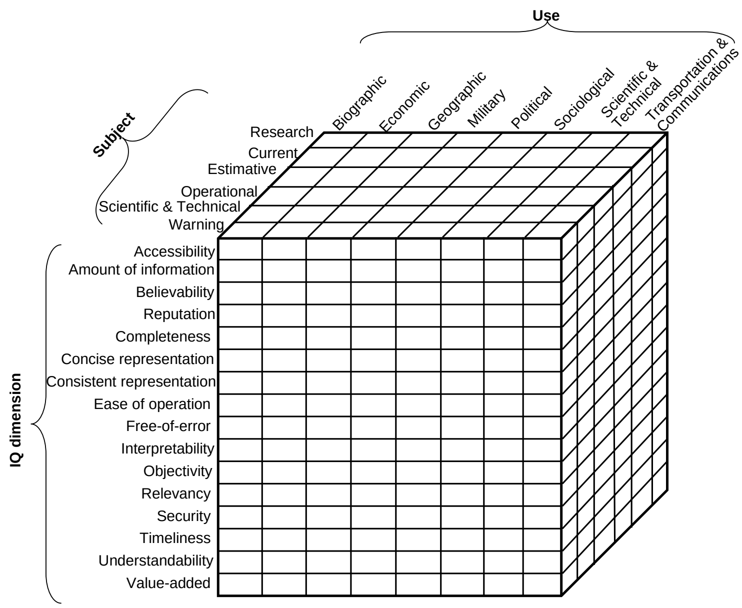

Fig. 4. Metrics for Intelligence Products and their Different Uses

For purposes of benchmarking and quality management, every intelligence product can be measured according to the quality metrics that correspond to the *subject* and the *use* of the product.

The specific objective of this research task is to verify, and help the IC to implement, the quality metrics system. Specifically, we need to:

- determine if the categorization mechanism is appropriate
- determine what quality dimensions are appropriate
- define each selected quality dimension and design instrument for measurement
- validate the instrument and verify the overall quality metrics system

4.2 Verifiability of Intelligence Products

Within the Intelligence Community (IC), it is desirable that the statements in a finished intelligence product are verifiable. In the ensuring discussion, we call these statements *conclusions*. A careful examination of this criterion reveals that there are three kinds of verifiability:

- Capability to show the information source and the analytical processes that lead to the conclusions. We call this capability *traceability*.

- Capability to show whether the information from the source reflects the real world situation, i.e., whether the information reflects an objective fact of the real world.
- Capability to show whether the information from the source reflects the actual perceptions about a real world situation. That is, suppose different entities have different perceptions about a fact, the information represents the true perception of each entity.

This study focuses one the *traceability* aspect of *verifiability*. The other two aspects rely on traceability. Because not all information is used for analysis, traceability will help to reduce cost of verifying information validity in terms of real world state or the perceived states.

4.3 Objectives and Plan

The primary objective of this study is to develop a methodology that can help improve the quality of the finished intelligence product. Specifically, the methodology will enable the IC to:

- include quality indicators in the finished intelligence products to provide an objective verifiable level of confidence about the conclusions; and
- allow consumers and analysts to trace the sources and the analytical processes that lead to the conclusions.

Appropriate quality indicators can help the consumers to avoid misinterpretation of the finished intelligence products. Traceability is an important mechanism for improving the quality of the finished intelligence products. With traceability, the IC can diagnose the entire production process to pinpoint the elements that cause intelligence failure, identify the bottlenecks that need further improvement, and optimize resources to improve the overall efficiency of intelligence production.

The investigation is driven by the following set of questions:

- Does the existing intelligence production process capture information that would enable traceability? Is it done systematically or in an ad-hoc fashion?
- Suppose the existing production process systematically captures information for traceability, is it done manually? How long would it take to trace from a conclusion to the sources of information that contribute to the conclusion?
- To what extent can computerization of the provision of traceability information help improve quality and productivity?
- How do we provide information to derive quality indicators in finished intelligence products?

We propose three tasks to address the issues raised by the questions:

- Investigate the existing process to identify the opportunities for quality improvement.

- Develop a methodology of enabling traceability and the provision of quality indicators in finished intelligence products.
- Verify the methodology through proof-of-concept prototyping.

Investigation of Existing Process

We propose to identify several cases to hand-simulate the existing process of intelligence production. The simulation results will allow us to identify the opportunities of modifying the process to improve the quality of finished intelligence products. It also provides the baseline information for benchmarking the effectiveness of the methodology that we will develop.

Several important questions to ask include:

- Given a finished intelligence product, does the consumer know the level of confidence in the conclusions?
- If the consumer asks for a confidence level, how much effort is involved to give a good answer?
- Can confidence level be quantified? Or what are the best ways of describing confidence level?
- If the consumer would like to know how a conclusion is derived, how much effort is involved in identifying the sources and analytical processes that lead to the conclusion?
- How is traceability information captured? explicitly and systematically, or the opposite?
- Is the traceability information easy to query and manipulate? By hand or using computer tools?

Methodology for Traceability and Quality Indicators

We propose to develop the methodology by adapting the TDQM framework and its relevant concepts and methods identified earlier to fit the specific needs of the IC.

The methodology will be applied to the set of cases identified in the preceding step to hand-simulate the improved process. The results should allow us to estimate the costs and benefits of the proposed changes to the existing process. We anticipate that the benefits will outweigh the costs, hence demonstrating the feasibility of applying the methodology to the IC.

Similar questions can be asked to determine the capability of traceability and providing quality indicators in the finished intelligence products.

Proof-of-concept Prototyping

We anticipate that certain components of the new process under the proposed methodology would be done manually because of lack of computerized tools. We will identify these components as targets for further enhancements through computerization.

One potential component is a tool for storing and manipulating traceability information. Our preliminary finding indicates that the IP-Map and the PolyGen model are appropriate for this task. The prototype will create a suite of tools for visualizing production process and processing data quality related information.

5 Conclusion

The IC has been following the Kent doctrine in its quality assurance practice. While helpful, the doctrine operates at a very high level and heavily relies on analyst's experience to ensure the quality of finished intelligence products. The existing evaluation framework only partially captures quality dimensions that concern the end users.

The IC can improve the quality of its intelligence products by incorporating an effective IQ framework. After reviewing the TDQM framework, we discussed the relevance of the framework with the IC's quality enhancement efforts. We also identified several concepts and methods useful to the IC. The proposed solution focuses on two areas: 1) developing comprehensive quality metrics for the IC; and 2) developing a methodology and a set of technologies to enable verifiability of intelligence products.

The initial focus will be on the traceability aspect of verifiability. With the proposed solution, we anticipate that intelligence production process will become more visible to the end consumers, and the quality of the products can be examined more easily. Consequently, the finished intelligence products will be more useful because the users can easily assess the confidence level of the produces and use the products more appropriately.

Future work will evaluate the proposed solution to identify areas for enhancement, e.g., incorporating entity resolution methods (Talburt et al, 2005; Wang and Madnick, 1989) to improve quality of information from multiple sources. Since we have only focused on the traceability aspect so far, future work will develop mechanisms to facilitate the improvement of the validity aspects.

Exercises

1. Briefly describe the TDQM framework and discuss how it can be applied to the IC as it tries to improve the quality of intelligence products.
2. Explain the concept of Verifiable Intelligence Product and discuss why it is important especially when complex processes and algorithims (e.g., data mining) are used to produce the final intelligence product.
3. Survey the literture and summarize your findings about techniques and products that enable traceability of information products.

Reference

Brei, W. S. (1996) Getting Intelligence Right: The Power of Logical Procedure. Occasional Paper #2, Joint Military Intelligence College (JMIC).

Buneman, P., Khanna, S., Tan, W.W. (2001) Why and Where: A Characterization of Data Provenance. In Jan Van den Bussche and Victor Vianu, editors, *International Conference on Database Theory*, pages 316-330. Springer, LNCS 1973.

Clauser, J. K. and Weir, S. M. (1975) Intelligence Research Methodology, An Introduction to Techniques and Procedures for Conducting Research in Defense Intelligence. Defense Intelligence School.

Deming, W. E. (1982) *Out of the Crisis*. MIT Press, Cambridge, MA.

Davis, J. (1992) "The Challenge of Opportunity Analysis. Intelligence Monograph", CSI 92-003U, Center for the Study of Intelligence.

Davis, J. (1995) Intelligence Changes in Analytic Tradecraft in CIA's Directorate of Intelligence. CIA Directorate of Intelligence.

Davis, J. (2002) The Sherman Kent Center for Intelligence Analysis. Vol. 1, No. 5, CIA: The Sherman Kent Center for Intelligence Analysis.

Directorate of Intelligence, Central Intelligence Agency (1997) A Compendium of Analytic Tradecrafts Notes. Vol. 1.

Garst, R. D. (1989) Components of Intelligence. In *A Handbook of Intelligence Analysis* (Ed, Garst, R. D.), Defense Intelligence College, Washington, D.C., pp. 1-32.

Harris, G. (1989) Evaluating Intelligence Evidence. In *A Handbook of Intelligence Analysis* (Ed, Garst, R. D.), Defense Intelligence College, Washington, DC, pp. 33-48.

Hulnick, A. S. (1988) Managing Intelligence Analysis: Strategies for Playing the End Game. *International Journal of Intelligence and CounterIntelligence,* **2**(3)**,** 321-343.

Jiang, L., Borgida, A., Topaloglou, T., Mylopoulos, J. (2007) Data Quality by Design: A Goal-Oriented Approach. The 12th International Conference on Information Quality, 249-263.

Jones, M. D. (1995) *The Thinkers Toolkit: 14 Powerful Techniques for Problem Solving*, Three Rivers Press, New York.

Juran, J., Godfrey, A.B. (1999) *Juran's Quality Handbook*. 5th Ed. McGraw-Hill, New York, NY.

Kahn, B.K., Strong, D.M., Wang, R.Y (2002) Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*, 45(4): 184-192.

Krizan, L. (1999) Intelligence Essentials for Everyone. Occasional Paper #6, Joint Military Intelligence College

Lee, Y., Strong, D. (2003-4) Knowing-why about data processes and data quality. *Journal of Management Information Systems 20*, 3, 13-39.

Lee, Y., Strong, D., Kahn, B. Wang, Y. (2002) AIMQ: a methodology for information quality assessment. Information & Management 40, 133-146.

Madnick, S. and Warng, R.Y. (1992) Introduction to Total Data Quality Management (TDQM) Research Program. TDQM-92-01, Total Data Quality Management Program, MIT Sloan School of Management.

Mathams, R. H. (1995) "The Intelligence Analyst's Notebook", In *Strategic Intelligence: Theory and Application* (Eds, Dearth, D. H. and Goodden, R. T.), JMITC, Washington, DC, pp. 77-96.

Mosier, D. (2005) Data/Information Quality in Intelligence Community. SAIC Whitepaper.

Pipino, L. Lee, Y., Wang. R. (2002) Data quality assessment. *Communications of the ACM 45*, 4, 211-218.

Shankaranarayan, G., Ziad, M. , Wang, R.Y. (2003) Managing data quality in dynamic decision environment: an information product approach. *Journal of Database Management*, **14**(4), 14-32.

Shankaranarayan, G., Wang, R.Y. (2007) IPMAP Research Status and Direction. The 12th International Conference on Information Quality, 500-517.

Strong, D. M., Lee, Y. W. and Wang, R. Y. (1997) Data Quality in Context. *Communications of the ACM,* **40**(5)**,** 103-110.

Talburt, J., Morgan, C., Talley, T. and Archer, K. (2005) Using Commercial Data Integration Technologies to Improve the Quality of Anonymous Entity Resolution in the Public Sector. *10th International Conference on Information Quality*, Cambridge, MA**,** 133-142.

Wang, R. Y. (1998) A Product Perspective on Total Data Quality Management. *Communications of the ACM,* **41**(2)**,** 58-65.

Wang, R. Y., Kong, H.B., Madnick, S.E. (1993) Data quality requirements analysis and modeling. In *Proceedings of the 9th International Conference of Data Engineering*, 670-677.

Wang, R. Y., Lee, Y. W., Pipino, L. L. and Strong, D. M. (1998) Manage Your Information as a Product. *Sloan Management Review,* **39**(4)**,** 95-105.

Wang, R.Y., Madnick, S.E. (1989) The inter-database instance identification problem in integrating autonomous systems. In Proceedings of the 5th International Conference on Data Engineering, 46-55.

Wang, R.Y., Madnick, S.E. 1990. A polygen model for heterogeneous database systems: the source tagging perspective. In *Proceedings of the 16th VLDB Conference*, Brisbane, Australia, 519-538.

Wang, Y. R. and Madnick, S. E. (1990) A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective. *the 16th International Conference on Very Large Data bases (VLDB)*, Brisbane, Australia**,** 519-538.

Wang, R. Y., Reddy, M., Kon, H. (1995) Toward quality data: an attribute-based approach. *Decision Support Systems*, **13**(3-4), 349-372.

Wang, R. Y. and Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems,* **12**(4)**,** 5-34.

Widom, J. (2005) Trio: A System for Integrated Management of Data, Accuracy, and Lineage. In *Proceedings of the Second Biennial Conference on Innovative Data Systems Research* (CIDR '05), Pacific Grove, California, January 2005