GEOSPATIAL DATA QUALITY: THE CONTENT MATURITY MODEL

John W. Strebeck
Emery T. Wilson
Thomas F. Creel, Ph.D.
Walter L. Lister
Paul X. Callahan
National Geospatial-Intelligence Agency (NGA)
7500 GEOINT Drive
Springfield, VA 22150-7500
John.W.Strebeck@nga.mil
Emery.T.Wilson@nga.mil
Thomas.F.Creel@nga.mil
Walter.L.Lister@nga.mil
Paul.X.Callahan@nga.mil

ABSTRACT

NGA is building a Content Maturity Model (CMM) to rate geospatial-intelligence (GEOINT) products, services and data. The CMM concept enables consumers of content to understand its quality and suitability for their mission, and also to have a conduit by which to provide feedback to NGA. The consumer rating capability will improve the feedback mechanism between consumers and producers of content, enabling NGA to provide rapid quality improvements and, subsequently, consumer application corrections. This initiative supports the NGA Strategic Objective on Content by creating and proliferating GEOINT content which is imperative to consumer requirements. CMM data quality components will be searchable, discoverable and provide indicators for consumers which informs them whether or not they are in possession of the best data available. The CMM also supports the Agency's transformation from a product-focused model to a data-focused provisioning solution. In this new data-centric environment, consumers will be serving or linking to many sources of GEOINT and a mechanism to receive and provide quality feedback will be critical. Consumers require indicators of data quality. The CMM is the means to benchmark geospatial data quality for consumer analysis as well as the data's own evolution.

KEYWORDS: data quality, data rating, metadata, quality metrics, NGA

THE DATA-RICH ENVIRONMENT: ROLE OF DATA QUALITY METRICS

In a data-rich, multi-provider environment where quality may vary, it is essential to have and maintain a two-way quality feedback mechanism that supports the consumer, provider, and broker. NGA's GEOINT stores will contain a full temporal spectrum of GEOINT data that may potentially contain disparate reporting to allow analysts the richest set of data sources and context. This new environment will allow NGA's consumers to leverage and ingest all types of content from numerous sources, including non-traditional sources, that continue to grow exponentially (Figure 1). This content includes, vector files, coverage data, gridded statistical data, photographs, video, compiled open source research reports, maps and text documents in varied formats.

The Content Maturity Model is structured to give both the consumer and NGA effective insight to data quality. As consumers begin to exercise creativity in application of content, they will need insight into the data's dimensional qualities in order to assemble "best of class" datasets for their mission-specific use. In a data-rich environment, routinely multiple datasets over the same object/event will be available and the consumer will need to differentiate and rank the datasets based on their unique applications and needs. In some instances, only a single dataset will be available for a given location, and the consumer will need to review its quality for the intended use, and the descriptive metrics derived from this process will serve as a baseline to improve such datasets.

Data stewards and spatial database managers will have the CMM as an additional tool to assist in managing

their content. In some cases, date of origin may be a driver for length of shelf life. However for other cases, very different attributes may provide better content management (e.g. the positional accuracy for roads in a given geographic area may tighten due to improved controlled imagery). Therefore, data stewards will use the CMM as a tool to review and evaluate the integrity of their dataset.

Figure 1. The GEOINT Data Store: a data-rich environment

EFFECTIVE DATA QUALITY METRICS: NEEDS OF THE CONSUMER

The core challenge in managing geospatial data quality is understanding the dimensional traits of spatio-temporal data and their relationships to consumers and consumers' requirements. The characterization of uncertainty about the accuracy of location is well known through Least Squares analysis and error propagation (e.g. Mikhail and Ackerman, 1976). However, uncertainty for abstract forms of characterization such as consistency and completeness, is much less known. Accordingly, a comprehensive evaluation of the parameters that affect Data Quality is the subject of much research (e.g. MacEachren, et al., 2005; Davis, 2012).

In light of the effort to build a framework that supports insight to data quality, the question arises as to how consumers of datasets use quality metadata. The literature does not indicate a significant body of related research. In a study that addressed the interpretation of data tags, it was deemed necessary for the tags to be blatant for users to get much value from them (Fisher, Chengalur-Smith, and Ballou, 2003). Another study found that when consumers are faced with a complex task, they give up on cognitive effort and do not utilize the insight from metadata tags (Price and Shanks, 2011). These consumer behaviors clearly impact decision accuracy, and increase random-error in final decisions.

While most quality metadata frameworks only reflect how the data was produced (e.g. Servigne, et al. 2006; USGS 1997), the addition of a metric that reflects consumer-centric views (Table 1) would emphasize issues such as the impact of uncertainties in the data (Goodchild, 2012). It is here that NGA seeks to foster an enhanced consumer

experience. Giving insight and a feedback mechanism that address data quality through the Content Maturity Model directly delivers that end. Providing consumers a metric that is simple to use will increase the likelihood of success and strengthen the probability to get the right content when needed, in the right format, and with context preserved.

Factor	Description
Descriptors	Simple descriptions of quality that are readily understood by non-expert users.
Impact of Uncertainties	Effects of uncertainties on specific uses of the data (from simple queries to complex analyses).
Tools	Tools to enable the user to determine the effects of quality on results.

Table 1. Factors in a consumer-centric view of data quality measures

THE MODEL

The Content Maturity Model promotes insight into quality in three dimensions (Figure 2); an NGA Analyst view, a Consumer or User view and a Fit-for-Use indicator. Core dynamics in this model are that consumers find data that are most suitable for their mission using the Fit-for-Use characterization, an NGA analyst review (NGA Analyst Rating), and peer reviews (User Rating).

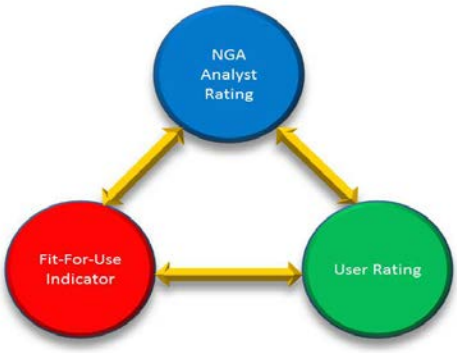

Figure 2. The Content Maturity Model and its components

Consumers will employ the User Rating to provide feedback on the value of data consumed, either at the time of data exploitation or after further evaluation. This value may also include a temporality aspect as content is needed, often within a specific time frame to satisfy intended applications. The feedback loop is intended to improve production processes and priorities, and also identify needed dataset enhancements. Through this process, the CMM fulfills its role as a data management tool.

Fit for Use Indicator

The Fit-for-Use indicator is assigned by the appropriate NGA-designated subject-matter expert (SME/ Analyst) for a dataset within his/her area of expertise. The utility of this rating is to facilitate a simple metric that describes the envisioned purpose of a given dataset. It is meant to support the consumer by defining what content should fit the criteria of a given mission. Note that quality and fit-for-use are not interdependent. For example, data or products can be of high quality but not fit for a particular use. A notional list of Fit-for-Use Indicators (categories) is provided in Table 2.

Fit-for-Use Indicators
Safety of Navigation
Strategic-Level Operations
Operational-Level Planning
Tactical-Level Operations
Not to be used for Safety of Navigation

Table 2. Fit-for-Use Indicators (*notional*)

User Rating

The User Rating (Figure 3) is the mechanism for consumers to rate data, enabling valuable feedback as to how the dataset is meeting their needs. The User Rating schema is the same across data domains and the identity of the user and date are included with the rating, which is visible to all of the data's consumers. If a wide variance exists for the User Ratings and the Analysts Ratings, an automatic notification is sent to the responsible content Analyst for review. To further understand the User Rating, the rater will self-identify their level of expertise as novice, intermediate or expert.

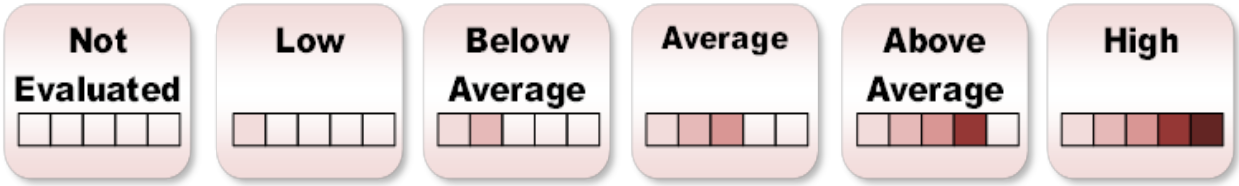

Figure 3. The User Rating Scale

NGA Analyst Rating

The NGA Analyst Rating is assigned by the SME analyst for a dataset within that analyst's area of expertise. The NGA Analyst Rating is created by evaluating the data against domain-specific criteria, all in the context of the utility of the dataset versus its intended use. Further, the NGA Analyst Rating is dynamic: it will evolve as current source information or insight into the data's potential for improvement becomes known, or as the data ages.

Most approaches to characterize components of quality are through frameworks, such as ISO 19114 (Table 3). In particular, ISO 19114 identifies quality elements critical to evaluation of spatial data. This ISO standard is incorporated in the criteria used within the CMM. However, for NGA-specific concerns, an additional criteria, Source Lineage, has been added as a category. Source lineage includes information about the data source, history, and production method that was not part of the current ISO standard. An Analyst Rating is provided for each quality element. The final rating is computed from these 6 elements and presented as an average. However, this average can be weighted. For instance, not all elements are of equal value to specific content. Some content may value completeness more than source lineage and thus a rating of 4 in source lineage would not carry as much weight in

the final average rating as a 4 in completeness. The star rating is a simplified quality rating that the consumer can quickly comprehend and is meant to ease the cognitive burden of making decisions of the best relevant data to support their mission.

Quality Elements	Description	Subelements
Completeness	Completeness is the presence or absence of features, their attributes and their relationships.	a. Commission - excess data in a dataset b. Omission - data absent from a dataset
Logical consistency	Logical consistency is the degree of adherence to logical rules of data structure, attribution and relationships (data structure can be conceptual, logical or physical).	a. Conceptual consistency - adherence to rules of the conceptual schema b. Domain consistency - adherence of values to the value domain c. Format consistency - degree to which data are stored in accordance with the physical structure of the dataset d. Topological consistency - correctness of the explicitly encoded topological characteristics of a dataset
Source Lineage	Source lineage includes information about the data source, history, production method and whether or not they are known, trusted, validated, etc.	
Positional accuracy	Positional accuracy is the accuracy of the position of a feature.	a. Absolute or external accuracy - closeness of reported coordinate values to values accepted as or being true b. Relative or internal accuracy - closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true c. Gridded data positional accuracy - closeness of gridded data position values to values accepted as or being true
Thematic accuracy	Thematic accuracy is the accuracy of quantitative attributes and the correctness of non-quantitative attributes and of the classifications of features and their relationships.	a. Classification correctness - comparison of the classes assigned to features or their attributes to a universe of discourse (ground truth or reference dataset) b. Non-quantitative attribute correctness - correctness of non-quantitative attributes c. Quantitative attribute correctness - accuracy of quantitative attributes
Temporal accuracy	Temporal accuracy is the accuracy of the accuracy of the temporal attributes.	a. Accuracy of time measurement b. Temporal Consistency/Correctness: correctness of ordered events or sequences c. Temporal Validity: validity of data with respect to time

Table 3. Quality Elements for evaluation of Spatial Data (from ISO 19114:2003)

The following (Table 4) is an example of a potential criteria matrix for an exemplary data type. The column headings reflect the criteria value and their relationship to the fitness-for-use of the given data.

Categories	★☆☆☆☆ Reference Only	★★☆☆☆ Use with Caution	★★★☆☆ Use with Confidence	★★★★☆ Use with Confidence	★★★★★ Use with Confidence
Positional Accuracy	Data Accuracy is non-specified	Data Accuracy complies with statistical measures with < 68% confidence	Data Accuracy complies with statistical measures with 68% confidence	Data Accuracy complies with statistical measures with 85% confidence	Data Accuracy complies with statistical measures with 95% confidence
Logical Consistency	No adherence to logical rules of data structure, attribution and relationships, Data Standards not employed	Adherence to logical rules of data structure and relationships, but without attribution. Data specified using proprietary formats	Adherence to logical rules of data structure, attribution and relationships. Data specified using proprietary formats	Adherence to logical rules of data structure, attribution and relationships, known data standards employed	Adherence to logical rules of data structure, attribution and relationships, Data standards conform to ISO standards
Source Lineage	Data source not known, validated or authoritative	Data source is known with no history of source authority or methodology	Data source and history of source is known, trusted collection methodology is known with incomplete metadata	Data source and history of source is known, trusted collection methodology is known with complete metadata	Data source and history of source is known, trusted collection methodology conforms to known standards with complete metadata
Completeness	Data collection is < 70% complete	Data collection is, >70% complete	Data collection is, >85% complete	Data collection is, >90% complete	Data collection is, >95% complete
Temporal Accuracy	Creation/revision/review < 10 and > 6 years	Creation/revision/review date < 6 and > 5 years	Creation/revision/review date < 5 and > 4 years	Creation/revision/review date < 4 and > 3 years	Creation/revision/review date < 3 years
Thematic Accuracy	Inaccurate quantitative attributes and unknown correctness of non-quantitative attributes and of the classifications of features and their relationships	Unknown or below 68% quantitative attributes, correctness of non-quantitative attributes, or the classifications of features and their relationships	Quantitative attributes, correctness of non-quantitative attributes and classifications of features and their relationships are known with 68% confidence	Quantitative attributes, correctness of non-quantitative attributes and classifications of features and their relationships are known with 85% confidence	Quantitative attributes, correctness of non-quantitative attributes and classifications of features and their relationships are known with 95% confidence

Table 4. Criteria Used to Determine Analyst Rating for Topographic Features

SUMMARY

The Content Maturity Model provides consumers a metric that is simple to use and apply. It will enhance the likelihood of successful integration of NGA-brokered content into customized products supporting specialized data needs. It is expected to improve the probability of consumer obtaining the right content when needed, in the right format, and with context preserved. However, this model also provides NGA with a critical quality feedback mechanism that is expected to enable very rapid response to consumer-related data issues.

REFERENCES

Davis, J., 2012. ESB Data Quality Standards, School of Library and Information Science, San Jose State University, 12 pp. http://slisapps.sjsu.edu/gss/ajax/showSheet.php?id=4943

Federal Geographic Data Committee, 2012. Business Plan for the Geospatial Platform—REDACTED. http://www.fgdc.gov/initiatives/resources/2012-09-12-geospatial-platform-business-plan-redacted-final.pdf

Fisher, C.W., I. Chengalur-Smith, and D.P. Ballou, 2003. The Impact of Experience and Time on the Use of Data Quality Information in Decision-making, *Information Systems Research, 14(2),* 170-188.

Goodchild, M.F. 2012. Beyond Metadata: Towards User-Centric Description of Data Quality, NGA Academic Research Program (Grant HM1582-07-1-2020), 6 pp.

MacEeachren, A. M., A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler, 2005. Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know, *Cartography and Geographic Information Science, 32*, No. 3, 139-160.

Mikhail, E.M. and F. Ackerman, 1976. *Observations and Least Squares*, University Press of America, New York, 497 pp.

Price, R., and G. Shanks, 2011. The Impact of Data Quality Tags on Decision-making Outcomes and Processes, *Journal of the Association for Information Systems 12(4)*, 323-346.

Servigne, S., Lesage, N., and Libourel, T. (2006). Quality components and metadata, In: R. Devillers and R. Jeansoulin (Eds.), *Fundamentals of spatial data quality*, pp. 179-208. http://liris.cnrs.fr/%7Esservign/12 Servigne-EN-vSS.pdf

USGS (1997). Spatial Data Transfer Standard (SDTS): Logical specifications. U.S. Geological survey, Virginia.

This paper approved by NGA for release (PA #14-150).