Assessing and Evaluating Standard Compliance with a State and Local Government GIS Metadata Profile in Large Geospatial Databases

Timothy Mulrooney
Dept. of Environmental, Earth and Geospatial Sciences
North Carolina Central University
Durham, NC, USA
e-mail: tmulroon@nccu.edu

***Abstract*— Under the supervision of the North Carolina Geographic Information Coordinating Council (NCGICC) and Statewide Mapping Advisory Committee (SMAC), a committee defined and developed a State and Local Government Metadata profile intended for use in North Carolina. This profile is based on the International Organization for Standardization (ISO) 191** standards. In addition to dictating best practices and conventions for existing metadata entries such as the Title, Publication Date and Use Constraints, this standard accounts for evolving technologies that did not exist when original metadata standards were first developed. While the rate at which geoinformation is created has exponentially increased, the time dedicated to cataloging and subsequently assessing and evaluating this metadata information remains nearly the same. In addition to educating the North Carolina Geographic Information Systems (GIS) community on this new standard, the research team is currently developing tools so GIS managers can gauge standard compliance more efficiently and proactively than in the past. In this short paper, the research team has begun using programming methods in which metadata entries from multiple layers in large geospatial databases can be assessed and evaluated. These methods will be tested using various quantitative methods, including the Technology Acceptance Model (TAM). This can provide insight into the various accuracies (horizontal, vertical, temporal, etc.) of layers which in turn can dictate future efforts. It can also be used to identify inconsistencies in metadata entries with an end goal of understanding misinterpretation of the profile so it can be improved in future incarnations.**

Keywords-GIS Metadata; Metadata; Metadata Profile; North Carolina State and Local Government Profile.

I. RATIONALE

A GIS serves as the tangible and intangible means by which information about spatially related phenomena can be created, stored, analyzed and rendered in the digital environment. In the North Carolina GIS community, GIS is used to represent transportation routes, elevation, delineate land ownership parcels, highlight patterns of crime and help make zoning decisions. The manner in which geospatial data is captured varies. Some methods include using a Global Positioning System (GPS) unit, extracting or improving existing GIS data, the use of an Unmanned Aerial Vehicle (UAV) or some other remote sensing platform, or creating data from an analog format via digitization. Regardless of the method, the resources (e.g., the computers, time and people dedicated to the process of collecting and creating geospatial data) are the most time-consuming portion of a GIS-related project [1]. As a result, the GIS community needs to ensure the quality of geospatial data created from these methods is captured and assessed in a systematic way.

Geospatial metadata serves as the formal framework to catalog descriptive, administrative and structural information about geospatial data. Geospatial metadata is inherently different from other forms of electronic metadata because each metadata file can be applied a spatial component that is not implicit with other forms of metadata. Given the capricious rate at which all forms of geo-information can be created, formal metadata serves as a lifeline between the tacit knowledge of the data creator and current and future generations of geospatial data consumers.

In the United States, the Federal Geographic Data Committee (FGDC) metadata standard, commonly referred to as the Content Standard for Digital Geospatial Metadata (CSDGM) allows for more than 400 individual metadata elements. The North Carolina GIS community has been proactive about understanding the importance of metadata. Under the supervision of the NCGICC and SMAC, a committee was tasked to develop a State and Local Government Metadata profile for geospatial data intended for use in North Carolina. This standard is based on the ISO 191** format and is an improvement over prior metadata standards to account for evolving technologies such as remotely sensed imagery, online services and ontologies. These were not considered when original metadata standards such as the CSDGM (formally known as *FGDC-STD-001-1998)* were first published. At this time, assessing and evaluating adherence to this standard for large spatial databases is an exhaustive process, as users must toggle through multiple levels of metadata records among multiple features a using a metadata editor. The goal of this paper is to propose a programmatic and faster assessment and evaluation alternative that can be used by GIS management to facilitate decision-making.

The rest of this paper is organized as follows. Section II describes the evolution of metadata. Section III describes the specific use and application of the North Carolina State metadata profile. Section IV addresses the how standard compliance is addressed. Section V discussed preliminary results. The acknowledgement and conclusions close the article.

ISBN: 978-1-61208-617-0

II. THE EVOLUTION OF METADATA SCIENCE AND ASSESSMENT

Although metadata's original use was simply as a means to catalog data, its storage and assessment has become a science in itself. The role of metadata assessment can be seen in a variety of different fields. An Electronic Metadata Record (EMR), for example, is a technology that is produced and edited when an electronic document is edited or created, such as a patient record or digital x-ray. Thus, the ease of storing, accessing and retrieving electronic metadata and files for medical data can help prevent litigation against malpractice lawsuits [2]. A complex statistical analysis was to retrieve biomedical articles from more than 4,800 journals to help support decision-making processes [3]. If properly maintained, metadata serves as a capable surrogate when querying scanned imagery or hard-copy information is not feasible and further validates in-situ decisions as they are reinforced by easily accessible support literature.

Early research and commentary on the concept of geospatial metadata has touted its value as an effective decision-making tool, regardless of its native format [4]. These formats include Hyper Text Markup Language (HTML), Extensible Markup Language (XML) along with its various ISO standards (19115, 19139), TXT (Text File), Geography Markup Language (GML) and Standard Generalized Markup Language (SGML), as well as proprietary formats. Methodology has explored the ability to integrate spatial metadata to a stand-alone database long before metadata was stored in a standardized format, as well as compiling statistics about metadata elements within the confines of specific software [5] [6].

The population of geospatial metadata is a monotonous process and subject to error, although research has explored the large-scale production of standards-based metadata in order to alleviate these issues [7][8]. Because of this, research maintains that human nature alone undermines the immediate and long-term goals of metadata for an organization and the GIS user community [9]. While the omission of one minor element would not degrade a layer's metadata or invalidate the geospatial data on which it is based, it may compromise quantitative data quality measures captured from which decisions can be made. More recently, feature level metadata has been able to capture data quality information, but is typically limited to quantitative measures of positional accuracy and qualitative information related to data lineage within eight of the more than 400 entries that comprise a complete FGDC-compliant metadata file [10] [11]. Even now, the population of these metadata elements is not fully automated and some entries must be done by a GIS data steward.

III. THE NORTH CAROLINA STATE AND LOCAL GOVERNMENT PROFILE

Geospatial metadata standards serve as a cohesive means by which organizations can define, store and more importantly share information about geospatial data. It defines the categories of information that needs to be stored, individual entries, or tags, of individual elements within these categories and the types of data (text, date, number) and their lengths that can be stored while expressing these tags. FGDC metadata is divided into 7 sections or divisions that transcend descriptive, administrative and structural components. They are: Identification Information, Data Quality Information, Spatial Data Organization Information, Spatial Reference Information, Entity and Attribute Information, Distribution Information, and Metadata Reference Information [12]

Within these high-level divisions, subdivisions and eventually individual metadata tags can be populated to catalog various forms of information about the GIS data layer. The hierarchy of these divisions and subdivisions are consistent with a standard. In addition to providing this structure, the FGDC also creates guidelines by dictating which metadata elements are to be populated. The FGDC requires seven metadata elements be populated for all GIS data. The FGDC also suggests that fifteen metadata elements be populated. These suggested and required elements are included in Table I below.

TABLE 1: REQUIRED AND SUGGESTED FGDC ELEMENTS

FGDC -Required Elements	***FGDC- Suggested Elements***	
Title Reference Date Language Topic Category Abstract Point of Contact Metadata Date	Dataset Responsible Party Geography Locations by Coordinates (X and Y) Data Character Set Spatial Resolution Distribution Format Spatial Representation Type Reference System Metadata Character Set	Lineage Statement Online Resource Metadata File Identifier Metadata Standard Name Metadata Standard Version Metadata Language

Organizations actively create content standards for new technologies and manners in which geospatial data are collected and stored. One such example is the FGDC content standard for Remotely Sensed Data. This includes two divisions germane to the equipment and methods such as platform name, sensor information and algorithm information used to capture the imagery, in addition to the seven existing aforementioned divisions [13]. Standards such as these and others must be increasingly flexible and updatable to account for the evolving technologies in which geospatial data can now be captured (crowdsourcing, Unmanned Aerial Vehicle, large scale geocoding), processed (new geostatistical and interpolation algorithms) and ultimately delivered (web map service, web feature service) to the GIS user community.

In recent years, the North Carolina SMAC has recognized most GIS data managers lack the time and resources necessary to learn and apply a metadata standard. To address the problem of missing or incomplete metadata records among state and local data publishers, the SMAC chartered an ad-hoc Metadata Committee in October 2012 to "recommend ways to expand and improve geospatial metadata in North Carolina that are efficient for the data producer and benefit data users in the discovery and application of geospatial data." The Metadata Committee submitted a draft of this profile, based on the ISO 19115 (for Geographic Information – Metadata: 2003), ISO 19115-1 (for Geographic Information – Metadata – Part 1: Fundamentals: 2014) and ISO 19119 (Geographic Information – Services: 2016) standards. After review and modification by SMAC and its standing committees, the most current version of this standard has been in effect since December 30, 2016 and is available through the NCOneMap portal [14].

Given seven required and fifteen recommended metadata elements are fairly ambiguous and less than ideal for many organizations whose data is integrated into the NCOneMap [15], the North Carolina state geospatial data portal, this profile provides explicit guidance on required/suggested metadata elements, wording for these elements, standardization of naming/date conventions and domain fields for topic categories for more than 75 metadata tags. A few examples of the rules for geospatial metadata include:

1. Publication Date is required and the format for Publication Date is YYYY-MM-DD or YYYYMMDD. If day is not known, use YYYY-MM and use YYYY if month is not known.
2. Abstract is required as a free text entry.
3. Status is required and only possible values are ‘historicalArchive’, ‘required’, ‘planned’, ‘onGoing’ ‘completed’, ‘underDevelopment’ and ‘obsolete’.
4. Topic Category is required and can be one of 23 possible values from domain table.
5. Online linkage is required to an URL address that provides access, preferably direct access, to the data

The following are additional examples of rules for Geospatial Services:

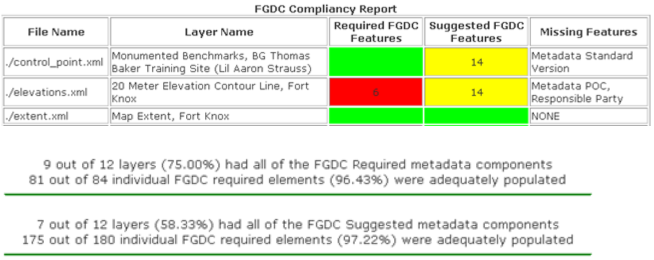
1. Metadata Scope code must be ‘service’.
2. Online Function code is required from domain of one of five possible values.

This richer metadata enables content consistency and improves the search and discovery of data through NCOneMap.

IV. ASSESSING STANDARD COMPLIANCE

Given the ever-increasing size of GIS data sets and the metadata requirements for each data layer, there needs to be a mechanism to assess the quality of these metadata not seen in previous generations or documented in existing literature. There also needs to be a means by which individual metadata entries adhere to predefined profiles and standards. Programming techniques and software packages have allowed users to assess information that would take a human days or perhaps weeks to do.

Open source solutions using Perl and R have been used to assess and evaluate metadata by traversing geospatial metadata stored in XML format as per FGDC requirements [16], resulting in quantitative metrics, graphs and reports regarding metadata compliance, as shown in Figure 1.



FGDC Compliancy Report

File Name	Layer Name	Required FGDC Features	Suggested FGDC Features	Missing Features
./control_point.xml	Monumented Benchmarks, BG Thomas Baker Training Site (Lil Aaron Strauss)	🟩	🟨 14	Metadata Standard Version
./elevations.xml	20 Meter Elevation Contour Line, Fort Knox	🟥 6	🟨 14	Metadata POC, Responsible Party
./extent.xml	Map Extent, Fort Knox	🟩	🟩	NONE

9 out of 12 layers (75.00%) had all of the FGDC Required metadata components
81 out of 84 individual FGDC required elements (96.43%) were adequately populated

7 out of 12 layers (58.33%) had all of the FGDC Suggested metadata components
175 out of 180 individual FGDC required elements (97.22%) were adequately populated

Figure 1: Sample of Metadata Compliance Report Generated Using Open Source Assessment Tool

As applied to the NC State and Local Government Profile, one major challenge exists. Primarily, geospatial data and metadata is typically software specific. While optimal open source solutions could be used to gleam information from metadata stored in XML using an appropriate xPath, these software-agnostic solutions are typically loosely-coupled and not intuitive to the average user. As a result of reliance on Esri products throughout the state, the Python programming language is being used to run this iteration of an assessment and evaluation tool before open source solutions are explored.

Using the NC State and Local Government Profile as a guideline, the research team has been developing tools for data managers to access and evaluate metadata entries. At the current time, metadata entries are written to CSV (Comma Separated Values). While doing this, string operations are run to ensure that required entries are populated, date entries comply with required conventions and domain entries match those in the domain table, all while agglomerating results and statistics at the database, layer (record) and tag (attribute) level. They can provide GIS managers with insight on non-compliant metadata entries to determine relationships between non-compliant entries and data steward or particular attributes that are continually non-compliant. The current working application of this code takes less than one minute to assess and evaluate 75 metadata elements for a GIS database containing 70 individual layers.

V. PRELIMINARY RESULTS

The TAM (Technology Acceptance Model) was used to assess and quantify the effectiveness of the open source metadata assessment tool. The TAM that we know of today was originally created as a means to universally quantify the effectiveness of technology by exploring relationships between the technology’s Perceived Ease of Use, Perceived Usefulness, Attitude Towards Using and the Intention to Further Use the technology [17]. Using Chronbach’s Alpha, Principal Components Analysis and Simple Linear Regression, associations can be found between these various components, as shown in Figure 2.

In this case, TAM has shown the potential effectiveness of this tool. However, H5 (Attitude Towards Using has a significant effect on Intention to Use) is not supported with 95% confidence. Possible reasons why this model is not supported is not a disconnect between these two concepts, but the actual implementation of technology given the role of the respondents. This survey used 50 respondents whose roles ranged from GIS technicians to GIS managers. GIS technicians working on few GIS data layers have little to no need for metadata assessment and therefore no intention to further use it. When enough GIS Managers have completed the assessment on which TAM is based, it will be run once again on this new tool to assess its effectiveness for a more germane usership.

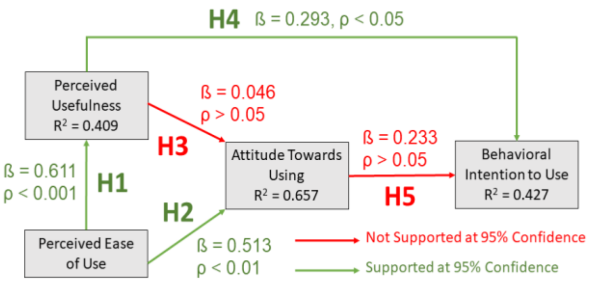

H4 ß = 0.293, ρ < 0.05
Perceived Usefulness R² = 0.409
ß = 0.046 ρ > 0.05
H3
ß = 0.233 ρ > 0.05
Attitude Towards Using R² = 0.657
Behavioral Intention to Use R² = 0.427
H5
ß = 0.611 ρ < 0.001
H1
H2
Perceived Ease of Use
ß = 0.513 ρ < 0.01
Not Supported at 95% Confidence
Supported at 95% Confidence

Figure 2: Regression Used to Test Research Hypotheses

VI. DISCUSSION

While a powerful and efficient tool, the programmatic assessment and evaluation of metadata entries still cannot altogether replace the human component. While these technologies can traverse metadata schema and extract tags to deem if they are complete, compliant or belong to a particular domain, it does not necessarily mean they are correct. QA/QC (Quality Assurance/Quality Control) techniques should be used to determine metadata quality across the entire dataset via ANSI (American National Standards Institute), ANSQ (American Society of Quality Control) or other institution-wide QA/QC procedures that best fit needs, resources and limitations.

VII. CONCLUSION

The increasing schism between the rate at which data are created and the efficiency at which the metadata are assessed serves as the impetus of this preliminary research. This paper looked to explore solutions to measure adherence to a state-level profile. Thus far, a programmatic solution using the Python programming language has been implemented. However, it is too early to tell how well these can be integrated into business processes at organizations such as the NCGICC. This ongoing research highlights the importance and need of programmatic approaches to the assessment and evaluation of metadata for large spatial datasets. This information can provide GIS Managers with already limited resources with the tools to make informed decisions that are not feasible with visual inspection or a qualitative knowledge of these increasingly large datasets.

ACKNOWLEDGMENT

The author wishes to thank the North Carolina Department of Transportation (NCDOT) for their generous support of this ongoing research.

REFERENCES

[1] K. Leiden, K. Laughery, J. Keller, J. French, J., W. Warwick and S. Wood, “A Review of Human Performance Models for the Prediction of Human Error,” Moffett Field, CA : National Aeronautics and Space Administration, 2001.

[2] T. McLean, L. Burton, C. Haller and P. McLean, “Electronic Medical Record Metadata: Uses and Liability,” Journal of the American College of Surgeons, vol. 206(3), pp. 405 – 411, 2008.

[3] T. Theodosiou, L. Angelis and A. Vakali. “Non-Linear Correlation of Content and Metadata Information Extracted From Biomedical Article Datasets,” Journal of Biomedical Informatics, vol. 41(1), pp. 202 – 216, 2008.

[4] D. Wong and C. Wu, “Spatial Metadata and GIS for Decision Support,” Proceedings of the Twenty-Ninth Hawaii International Conference, vol. 3 (3 – 6), pp. 557 – 566, 2006.

[5] D. Lanter, “A Lineage Meta-Database Approach Towards Spatial Analytic Database Optimization,” Cartography and Geographic Information Systems, vol. 20(2), pp. 112-121, 1993.

[6] D. Lanter, “The Contribution of ARC/INFO's Log File to Metadata Analysis of GIS Data Processing,” Proceedings of the Fourteenth Annual ESRI User Conference, Palm Springs, California, 1994.

[7] G. Giuliani, Y. Guigoz, P. Lacroix, N. Ray and A. Lehmann, “Facilitating the production of ISO-compliant metadata of geospatial datasets,” International Journal of Applied Earth Observation and Geoinformation, vol. 44, 23-243.

[8] S. Trilles, L. Diaz and J. Huerta, “Approach to facilitating a geospatial data and metadata publication using a standard geoservice,” International Journal of Geo-Information, vol. 6(5), pp 126.

[9] C. Doctorow. *Metacrap: Putting the Torch to Seven Straw-Men of the Meta-Utopia*. [online]. Available from http://www.well.com/~doctorow/metacrap.htm. [retrieved February 2018].

[10] L. Qiu, G. Lingling, H. Feng and T. Yong, “A unified metadata information management framework for the digital city,” Proceedings of IEEE’s Geoscience and Remote Sensing Symposium, pp. 4422–4424, 2004

[11] R. Devillers, Y. Bédard, and R. Jeansoulin, “Multidimensional management of geospatial data quality information for its dynamic use within Geographical Information Systems,” Photogrammetric Engineering and Remote Sensing, vol. 71(2), pp. 205–215, 2005.

[12] Federal Geographic Data Committee (FGDC), “Content Standard for Digital Geospatial Metadata Workbook,” Washington D.C.: Federal Geographic Data Committee, 2000.

[13] Federal Geographic Data Committee (FGDC), “Content Standard for Digital Metadata: Extensions for Remote Sensing Data,” Washington D.C.: Federal Geographic Data Committee, 2002.

[14] North Carolina Geographic Information Coordinating Council (NCGICC), *North Carolina State and Local Government Metadata Profile for Geospatial Data and Services* [online]. Available from http://www.nconemap.gov/DiscoverGetData/Metadata.aspx#iso. [retrieved February 2018]

[15] North Carolina Geographic Information Coordinating Council (NCGICC). *North Carolina OneMap* [online]. Available from http://www.nconemap.gov. [retrieved February 2018].

[16] T. Mulrooney, “Turning Data into Information: Assessing and Reporting GIS Metadata Integrity Using Integrated Computing Technologies,” Greensboro, North Carolina: University of North Carolina, Greensboro, 2009.

[17] F. Davis, “Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology,” MIS Quarterly, vol. 13(3), pp. 319-340, 1989.