

Item Level Metadata for Volunteered Geographic Information

Eric B. Wolf, University of Colorado at Boulder
ebwolf@colorado.edu

Metadata serves an important role in describing the quality and establishing the credibility of geographic data. Metadata for geographic information has been formalized in a manner not common in most disciplines. Almost two decades ago, the Federal Geographic Data Committee (FGDC) adopted the Content Standard for Digital Geospatial Metadata (CSDGM) and it's been almost a decade since the the International Standards Organization (ISO) 19115:2003 schema for metadata for geographic information was adopted. Both of these standards explicitly specify how the quality of data should be described. However, these standards and their implementation are based on a data production methodology that did not predict volunteered geographic information (VGI).

Geographic information has traditionally been produced through formal surveys relying on closely managed processes, resulting in relatively uniform data quality. VGI may be produced by a number of contributors with varying levels of expertise, utilizing a variety of collection methods. The level of expertise and the collection method both impact the quality of data collected.

The presented research explores the structure of metadata in the National Hydrological Database (NHD) managed by the US Geological Survey (USGS). The NHD is maintained through a network of State-appointed Stewards who manage VGI updates to the NHD for their state. My research began by interviewing a sample of these Stewards to understand how the NHD Stewardship process is practiced. Two important results came from these interviews.

First, a typology of update practices was established. Some Stewards accepted updates from a wide variety of contributors. Some Stewards were primarily engaged in aggregating large database updates from other authoritative land managers such as Federal agencies and Regional water management districts. Other Stewards only produced updates through closely controlled, internal data production methods. These three practices correlated with population density and the amount of land cover occupied by water. Densely occupied states with a large percent of surface water tended to use the closely controlled method. Sparsely occupied states with very little surface water tended to use the aggregated method. Intermediate states relied on a more flexible update regime.

Second, the interviews clearly established that the current metadata structure for the NHD fails to capture the heterogeneity of data quality in updates submitted by Stewards. Metadata for the NHD is stored at the Digital Update Unit (DUU) level which corresponds to a single update submission by a Steward. Almost all Stewards stated that they regularly combined contributions from multiple contributors into single DUUs. Any variation in data quality resulting from differences in data collection methods used by contributors was not documented in the metadata.

Based on the results of the interviews, my research proposes a new *item level* structure for metadata. Following a concept suggested by the ISO 19115:2003, the item level varies based on how the attributes of the data being described varies. To demonstrate how item level metadata can be realized in the NHD, data quality metrics were created for positional accuracy, currentness and completeness for three NHD subbasins and stored at the reach level. An

aggregation tool was also created to facilitate data evaluation based on the resulting metrics.

Another important trend was noted during the interviews with Stewards. Any effort to improve metadata must not increase the level of effort necessary to produce metadata. The purpose of using calculated metric for data quality is to automate the metadata production, avoiding the burden on the data producer to write data quality reports in the metadata. The aggregation tool reduces the burden on the data consumer by providing a means to explore data quality at the desired aggregation level rather than a level determined by the data producer.