Insights | Global Geospatial Data Curation: Monitoring Data’s Value

Kevin Howald, Geospatial Collaboration Manager
Harris Corporation
October 2018
Global Geospatial Data Curation:
Monitoring Data’s Value

WHAT IS DATA CURATION?

Simply put, data curation is the active management of data throughout its useful life cycle. Curating data involves combining and organizing data across various sources, identifying and presenting the data’s value, and monitoring its usefulness over time to know if data is fit for use now or if it needs an update.

For geospatial big data environments, data curation includes monitoring:
- Quality measures of the data
- Data attributes like date and time of collection, location coordinates, type of information, or any specific classifications in the metadata
- Metadata tagging accuracy
- Data provenance or origin information like where it came from, what the purpose of the collection was, who maintains the data, or where the data has moved over time
- Data ownership
- Data transformations, modifications, or adjustments made to the data, perhaps to standardize it before integration and track its provenance over time
- Data currency or the availability of newer information
- Overall health of the data system

About the Author
Kevin Howald | Geospatial Collaboration Manager

Kevin Howald is geospatial collaboration manager for the Geospatial Solutions business unit within Harris’ Space and Intelligence Systems segment. With nearly 30 years of experience in geospatial data management, Howald has participated in a wide range of activities to support management, access, and exploitation of geospatial data and imagery. He led software development teams in building tools that evaluate, label, and extract data from imagery. He has also developed geographic information system (GIS) data analysis models, digital cartographic techniques, and assisted in database requirement development.

Prior to Harris, he served as regional manager for LizardTech, helping deliver solutions for massive, high-resolution geospatial data. He also served as a project manager for T.Kartor Sweden, providing technical guidance for GIS projects and customer experience, and he was a senior scientist for MDA Federal, where he supported programs involving imaging data, GIS, spatial analysis, data integration, and mapping.

Howald is currently a member of the United States Geospatial Intelligence Foundation Certification Governance Board and the Open Street Map Foundation. He has a master of arts degree in geography from the University of Oklahoma and a bachelor of arts degree from the University of Toledo.

Kevin.Howald@harris.com | (321) 984 6987

Harris’ curation tools are leading the way to geospatial big data stewardship. We are creating technologies that automate and enhance value assessments of existing data by auto-tagging data, monitoring attributes, assessing the quality and relevance, and identifying data gaps or insufficiencies across large areas. Our curation tool reduces duplication of effort in data collection and eases the burden of assessing mission-critical data.

What is the Real Value of Monitoring Data?

Curating data in a big data environment enables access to available data across multiple systems, reduces duplication of effort for data creation, and enhances the long-term value of existing data by making it available for other uses. These, in turn, significantly reduce the time required for processing and delivery of products compared to traditional methods using disparate systems and tasking analysts to locate mission-related data across the various systems.

How Does It Work?

Modern systems integrate source data that can add value, harvest the content, validate it, match it up to existing data, incorporate it into the system, and then monitor its usefulness. This is accomplished by engineered data management systems that perform automated workflows.

By incorporating modern data management tools, information is more quickly available to end users, so they can focus their attention on analysis tasks and develop mission-focused products with confidence.

Delivering Confidence
Effective data curation reduces labor hours spent locating fit-for-use geospatial data, enabling analysts to create high-confidence products more efficiently.
Global Geospatial Data Curation: Monitoring Data’s Value

Curation Fundamentals
Modern data management employs architecture and software tools that automate as much of the manual tasks as possible to ensure that data integrates smoothly, is processed quickly and made available when it is needed, and has a clearly identifiable value.

The curation need arises when an organization has multiple data sources within and outside the organization that supply potentially meaningful data for its missions. These sources each have their own management structures, querying interfaces, and unique access points. To successfully curate data and evaluate its potential, we need a way to view the data as a whole or integrate it.

Federation Integration
Data silos are a thing of the past. Modern data management systems integrate and standardize content to ensure the best, validated information is available and accessible to all users. With new technology and techniques for federation and integration, separate data systems can be accessed using a single entry point to allow equal access to users systemwide without needing to run multiple queries in multiple systems. This can be performed by central repositories or federated access systems.

A central repository warehouse merges data from a variety of sources into a master reference database. It is generally separate from operational systems, so it does not slow down operations as it ingests new data, and the native data (without modifications) is kept for tracking provenance and reference purposes. Central repository warehouses enhance the speed of data retrieval, improve processing power capabilities, and provide rapid access to information.

One example is Harris’ 1Object 1Time (1O1T) data warehouse, a PostgreSQL-based geospatial data framework that houses worldwide layers of topographic content. The 1O1T tools service-enabled content for high-speed, simultaneous access by multiple users on any distributed network or cloud.

Data federation keeps data where it resides in the individual silos or native locations, with rapid access to the information provided by indexing the available data at each source and creating a virtual database to run queries and locate data. This type of data integration causes less disruption, requires less infrastructure and manpower to implement, and provides a “window” into the full array of information available across multiple systems. It is a cost-effective option that is appealing to organizations with multiple data stores and a growing number of data sources.

By keeping data where it was originally stored, federated data access systems reduce storage requirements and ensure the most current data is available to multiple users across the organization.

For example, One connection has been made to any Open Geospatial Consortium (OGC) standardized data layer, that can be easily discovered and retrieved. After indexing, the data will be accessible at much faster speeds.

Data Labeling and Matching
Curation and modernized data management tools begin with data labeling. Data labeling is how we assign meaning to raw data; it describes the essence of what the data is providing to enable identification, categorization, and quick assessment of the information it contains.

Database systems use this labeling to sort the data and incorporate it into results for queries. Without the labels, search queries do not know what to do with the data, and it is ignored.

With the massive influx of data today, much of the available data has not been labeled or the labels do not have any meaning to the data systems. Harris’ tools align the datasets and automate basic labeling to ensure successful integration.

Historic datasets, data from multi-modality sensors, and data from emerging commercial sources need to be evaluated and matched up before integration to ensure integrity, viability, and accuracy. By connecting and integrating these various datasets and providing an easy-to-use access point, Harris is making the search and discovery task much easier for today’s analyst.

Analysts also need tools that match up the data, so they can compare or combine datasets. This involves interpreting specific geospatial data points to understand where the data overlaps. Harris’ feature-matching engine automates feature-to-feature comparison of vector geometry based on proximity, orientation, and attribution. This tool identifies which source is the best fit before the information is incorporated into the master dataset. In areas where potential conflicts exist, automated conflation tools address and resolve the conflict before the data is inserted.

Designed for Speed
The design of the data system is crucial to the accessibility and speed of the results. Harris is designing systems that can scale in size easily and incorporate new database functions to expand capabilities as technology advances.

The PostgreSQL design of our data store systems processes data at faster speeds than industry-standard geographic information system (GIS) tools. This is accomplished by how efficiently the system relates to feature sets and data tables. Our server-side, multi-threading design stores and processes the data, and then uploads the information to operations. This saves time on uploading and keeps source data intact.

Federated systems enable analysts to tap into multiple source datasets for immediate search and discovery while data systems are being continually updated. The system indexing essentially gives the analyst access to the information without slowing down the processes for the operational system or the search and discovery.

Scalable Systems for Exploitation
The best data curation systems are designed for today, but also are engineered for tomorrow’s ever-changing needs. Systems must be scalable to meet new needs as they arise. They need to be inclusive to allow integration of new sources, and they need to be flexible to operate on-premise or in shared cloud environments.

Systems must be able to handle the influx of increased data from the growing number of data-collecting sensors on the ground, in the air, and in space.
Global Geospatial Data Curation: Monitoring Data’s Value

**Content Exploration**

Data structures that federate or combine datasets provide single-access points from which to conveniently explore more content across multiple systems on premise or via cloud storage. While the fast search and discovery capability provided by geospatial big data sources is helpful, analysts may still be required to spend time tediously filtering out what is unimportant from the resulting large amount of data. So, the quality of results—not just speed—is critical.

Further automation and enhanced technologies can be implemented to pare down the data and provide higher quality results for the analyst.

**Content Assessment**

Evaluating the quality of data to be integrated or federated is essential to ensure the data transformation, normalization, or error detection and correction have been completed. This allows features to match properly and attributes to meet specifications. Ultimately, it helps analysts get accurate information.

Harris’ suite of active content assessment tools generates objective metrics across various sources, including feature area and volume, data quality dimensions, and comparison between sources. These tools ensure accuracy in matching the data so that all information fits within the shared environment.

An important part of curation during the data integration, or matching phase, is to maintain the metadata history or origination information. Harris’ assessment tools manage the metadata as a curator might do with artifacts at a museum, maintaining where the data came from, how it was collected, and what modifications may have been made upon its insertion into the system. This metadata history, along with data timeliness, helps to qualify data for mission use. Analysts gain confidence in their decisions knowing information is current, or they can use the metadata to assess the credibility of dated information.

**Reliability and Validation Testing**

Enterprise-wide quality management tools provide a reliable, accurate way to automate data validation and correction of geospatial vector content, increasing the quality of output from queries. Harris’ tools check data against a rigorous set of business rules to gauge the integrity of the data for analysts’ use.

**Evaluating Data Usefulness**

Much like the socialized product rating tools on an ecommerce website, it is often valuable to understand how others used products and or data sources. Harris has developed tools that allow for inhouse ranking or rating of data sources and data products that allow analysts to share feedback. This is helpful to inform other analysts as to how reliable a source may be or to understand if data attributes are not accurate before the data is selected for a mission.

This simple application can be integrated into an agency’s in-house rating system or content maturity hierarchy to ensure analysts get reliable information for their analyses.

**SUMMARY**

Successful modern geospatial big data management employs architecture and software tools that automate as much of the mundane tasks as possible to facilitate smooth data integration, rapid processing, and the accessibility and availability of data for exploitation. Harris is utilizing a combination of domain expertise, engineering architecture excellence, and data science to ensure the best information is available to customers when and where needed.

By incorporating modern data management tools, information is more quickly available to analysts, so they can focus their attention on the analysis and develop mission-focused products with confidence.