Paper for Consideration by S-100WG6

Add the possibility to store feature oriented discrete coverage in HDF5

Submitted by:	NOAA (USA), BSH (Germany)
Executive Summary:	Add the possibility to store feature oriented discrete coverage in Part 10c of S-100 Specification.
Related Documents:	S-100 Specification Edition 4.0.0 ISO 19129 S102PT6(2020) Germany/BSH S-102 extension as proposal to implement source metadata
Related Projects:	S-102 Product Specification

Introduction / Background

BSH has undertaken the task to enable the creation of S-102 datasets and has identified an area for enhancement of the S-100 specification to provide more possibilities to transport spatial information in the HDF5 format. Currently, when using grids, the S-100 only provides the ability to store individual information for grid tiles. Grid cell differentiated information transfer is currently not possible.
We consider future S-102 data products as regular tiles that can be inserted into the tiling scheme of SOLAS ENCs. We consider it necessary to link data sets in a product specification with additional feature information. Of what kind this additional information is does not matter for this technical proposal. However, for the sake of simplicity, it may be necessary to support some statements with examples. The examples here are oriented to S-102 and refer to information on individual survey datasets (features with differentiated attributes). We consider it important to point out that these are only examples. This technical proposal has a general and reusable character.
A possibility to link datasets with additional information is provided by the sections of Part 3 clause 7.4 and Part 8 clause 5.3 and Appendix 8-E of the S-100 Edition 4.0.0. The possibility of discrete grid coverages described there allows to store an ID as a reference to a dataset with information about the respective feature.

Analysis/Discussion

In the current description, the ID of the feature is stored at the discrete grid coverage and refers to an external GML file. The GML file has significant disadvantages compared to the technical solution described in the rest of the paper:

- Data consistency is not technically guaranteed
- Potential source of errors when creating the exchange set
- Data size unnecessarily large
- Complexity of the GML geometries can lead to display problems

The GML file contains the geometry of the feature and, if necessary, additional information about the feature. We consider the encoding of references to the GML file described in clause 10c-13 of S-100 to be impractical for use with raster geometries. We will therefore describe a technical implementation in the following paragraphs using an example from S-102.

Supersede GML

As a very first point in the technical implementation of discrete grid coverages, we see several problems with the GML file. The GML file is defined as an external file besides the actual dataset. Furthermore, GML is a text format with many additional descriptive expressions (tags) derived from the XML format. Also the specification of a vector geometry is mandatory in GML. These characteristics of the GML format lead to the following problems:

1. External file
   The use of an external file involves a number of data consistency risks. A clear affiliation between the actual dataset and the GML file must be ensured. This must not be technically eliminated under any circumstances. This cannot be permanently ensured by means of a naming convention. File names can be changed accidentally by human interaction, for example. The references described in clause 10c-13 of the S-100 would therefore be invalid.
2. Representation of the geometry of raster features

The mapping of raster geometries into vector geometries is always very complex due to the rectangular characteristics of the raster geometries. For an accurate mapping, the vector geometry must map the step effects of the raster geometry. This can only be achieved by a large number of vertices.
Furthermore, both raster and vector geometries can contain a large number of exclaves and enclaves. A graphical representation of the named effects can be derived from Figure 1. This phenomenon does not make the mapping of vector geometries in a GML file with nested tags any easier. Mapping vector geometries in textual format thus increases the risk of incorrect interpretation while reading and writing the geometries.

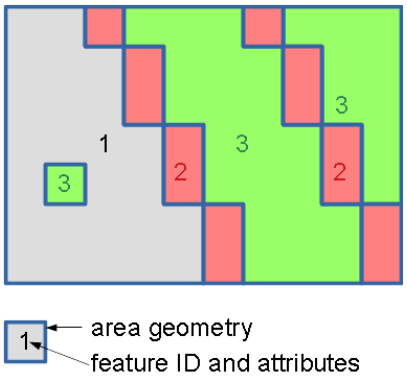


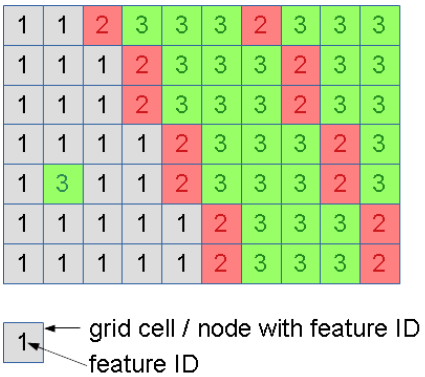
Figure 1: Raster geometries with exclaves, enclaves and step effects

3. GML file size
   The complex geometries mentioned in the previous paragraph leads to another problem in the file size of the GML file. Complex geometries require much more storage space in a textual representation than in a binary or raster representation. Furthermore, textual geometry descriptions are a bad basis for compression techniques because the coordinates occur in an almost random combination and have few commonalities.
   Also not conducive to file size reduction is the general textual representation of a GML file. Special attention should be paid to the use of tags. These are always specified twice (opening and closing) and additionally repeated for each feature.

Due to the problems described above, we recommend not using a GML file for feature information on raster geometries. In the following sections, we will present a more efficiently solution for mapping and managing feature information for raster geometries in HDF5 format.

Mapping in HDF5

The alternative proposal to using an external GML file is to use a separate regular grid and an attribute table within the HDF5 file. The grid corresponds in its extent and orientation to the actual data of the respective product specification. However, this does not represent continuous grid coverage, but rather feature-oriented discrete coverage as described in S-100 Edition 4.0.0 Part 3 clause 7.4 as a feature model. The geometry of the features are mapped in the form of a grid and assigned a feature ID. The information of the features is mapped in an attribute table and labeled with feature ID. The linking of the geometry with the information of the features is done using the already mentioned unique feature ID. A graphical representation of the interaction of grid and attribute table can be derived from Figure 2.



feature ID	attribute_1	attribute_2	attribute_3
1	17	5.3	text value A
2	6	14.7	text value B
3	10	5.3	text value C

Figure 2: Feature grid with attribute table

Feature ID

The feature ID can be expressed in various forms. It can be represented in strings and integers. However, NOAA/BSH recommend not to use strings. Strings require more storage space and are difficult to increment automatically.

NOAA/BSH recommend the use of natural numbers for the feature ID. Thus, an automatically generated random or sequential unique number can be generated. Likewise, an ID from an external data management system can also be used as the feature ID. In object-relational databases, for example, natural numbers are used to identify records. By using natural numbers, it would thus be possible for data producers to create a permanent and traceable reference to their own data management systems.

The fill value of the feature ID, if required, would then also be derived from the range of natural numbers and set to zero. Typically, zero is not used as a value for identifying records even in object-relational database systems (1-based index). Thus, the fixed interval of possible feature IDs has an open end in the range of $1 \geq X < \infty$.

The proposed data type would be an unsigned integer in HDF5 format.

New dataCodingFormat

The current variants of the dataCodingFormat (see S-100 Ed 4.0.0 Table 10c-10) do not provide for the possibility of coding the combination of grid and attribute table described here. Theoretically, it is possible to implement the mapping in the respective product specifications. However, in the interests of standardization and reusability, it makes sense to include this new coding variant in the S-100. The new dataCodingFormat is based on the dataCodingFormat=2 (Regular Grid). In addition, a further dataset of type Compound must be added in table 10c-18 of the S-100 Edition 4.0.0 to enable the mapping of the attribute table for the new dataCodingFormat. The new dataset has to be provided with a “required” rule according to the other datasets.

Conclusions

The new method presented here for encoding feature information in HDF5 format is an addition to the previous capabilities. It enables the simple, efficient linking of raster geometries and feature information without a lossy or complex conversion to vector geometries. This makes it possible to store feature information also on grid cells and not only on grid tiles. By not using an external GML file and mapping the feature information within the HDF5 file, the consistency of the data to each other is ensured. At the same time, the data size is reduced by the compression within the HDF5 format and removing the need of a the textual storage GML file. The transport of the data to the end user is thus more efficient. The use of a feature ID in the format presented here enables the data producer to establish a unique link between the S-1XX product and the data producer's data management system. A fast and efficient linking of the product data to the respective data origin is thus possible.

Recommendations

We recommend extending S-100 with the ability to link raster data to feature geometries and feature information presented here as an alternative to the current capabilities in HDF5 format.

To supplement this paper, we provide datasets. The sample dataset contains a general implementation of the proposal using a simplified geometry of the figures in this paper. The test dataset contains an implementation using a real dataset from the S-102.

Furthermore, we recommend the textual elaboration of the corresponding changes to the S-100 in a small group (break-out session) after the proposal be fundamentally accepted.

Justification and Impacts

We see the advantages of the solution presented here in the enrichment of the actual data of the respective product specification with additional data. Thus, the consistency between different S-1xx product specifications is improved. The additional data can convey further safety-relevant information to the end user, thus ensuring the greatest possible safety for the ship. In addition, the error-proneness and complexity of mapping raster geometries into vector geometries is eliminated. The reduced data size compared to the GML file increases the efficiency and speed of transporting the data to the end user, thus reducing transmission costs.

Action Required of S-100WG

The S-100WG is invited to:

a. Endorse the extension of the scope of S-100 Edition 5.0.0 Part 10c to include a feature oriented discrete coverage in HDF5.

b. Identify issues, which should be addressed in the course of the review and revision, in addition to those listed in this paper.

c. Take other actions as appropriate.