Paper for Consideration by S-100 WG 5

Validation checks for HDF5 and GML Data Products

Submitted by:	Raphael Malyankar on behalf of the S-100 WG Chair
Executive Summary:	Presents a draft specification of validation checks for S-111 (Surface Currents). Also presents a subset of validation checks for S-100 GML datasets from S-122/S-123/S-127. These check specifications are expected to evolve into generalized validation checks for S-100 HDF5 (Part 10c) and GML (Part 10b) formats respectively.
Related Documents:	--
Related Projects:	S-111; S-127; Other product specifications using HDF5 and GML data formats; S-98

Introduction / Background

A set of validation checks has been developed for S-111 (Surface Currents) HDF5 datasets. This set builds on the S-57 ENC validation checks, as updated for S-101, and their generalization for the ISO 8211 format at S-100 WG4 (S-100 WG4-4.5). This paper presents these checks as a prototype which can be generalized into a set of core checks for the S-100 HDF5 format.

Similarly, some product specifications which use the S-100 GML format have defined validation checks. S-122 (Marine Protected Areas), S-123 (Marine Radio Services) and S-127 (Marine Traffic Management) all include validation checks. These validation checks can be generalized into a set of core checks for the S-100 GML format.

References

S-100WG4-4.5 S-100 Validation checks.

S-58 ENC Validation checks, edition 6.1.0, September 2018.

S-98 Annex B Validation checks. S-98 draft 0.4, in the S-98 documents package, S-100 WG5 5.1B.

Discussion/Analysis

Annex A accompanying this paper defines concepts, terms, syntax, and operations used in the check descriptions. Annex A is based on Sections 1 and 2 of S-58, customized for the HDF5 and GML formats. Annex B is a spreadsheet containing checks for the HDF5 and GML formats (in two separate sheets repectively). The HDF5 set is the full set of checks for S-111; the GML set is a partial set which omits topological integrity checks. The checks are based on the S-101/S-57 validation tests and the proposed general set for S-100 ISO 8211 format in S-100WG4-4.5.

The Annexes differ from the documents on which they are based. The divergences are:

- The HDF5 format in S-100 Edition 4.0.0 does not cover vector spatial objects, and therefore the material in S-58 and its successors in S-122/123/127 concerning (vector) geometry and spatial operators should be omitted for S-111 and other product specifications using only the HDF5 format.
- The HDF5 format is more structured than either the ISO 8211 or GML format, and some checks depend on the existence of certain structural elements or metadata attributes, or on the values of certain metadata attributes. The checks have therefore been divided into phases which can be executed only if the required elements or metadata attributes are present and valid. Execution of later phases is controlled by means of semaphores which indicate whether the necessary items are missing or invalid.
- Some HDF5 checks depend on the success of other checks or on the values of metadata attributes, or require initialization of context from metadata attributes. Dependencies on other checks are indicated in a “Prerequisite checks” column, and dependencies on attribute values and requirements for context evaluation are indicated in a “Context test or initialization” column.
- For the S-100 GML format, update dataset formats as described in S-122, S-123, and S-127 are based on the principle of replacing “whole objects,” and therefore certain checks cannot be

run on update datasets alone but have to be executed after the update has been applied. This is the reason for distinguishing between base and update applicability for the base datasets.

The meanings of the various columns in the check specifications are explained in Annex A accompanying this paper.

The flow of check processing for the HDF5 set of checks is also described in Annex A. At present, the flow diagram is applicable only to HDF5 checks, but consideration will be given to developing a similar diagram for the GML format.

Consideration will also be given to describing test methods for the GML checks in particular, some of which can be checked by ordinary XML validation tools and others by means of Schematron rules (there are Schematron validation files included in the S-127 and S-98 schema packages). Note that it is possible to convert Schematron rules into XSLT templates.

Consideration will also be given to adding topological integrity checks to the GML generalized set, probably by adapting the relevant ISO 8211 checks or referencing them.

The S-111 validation checks include feedback received from developers working on the S-111 and S-104 product specifications and have also been circulated to the S-102 project specification development leads.

Annexes A and B accompanying this paper are illustrative outlines and may contain discrepancies. Full technical reviews of their contents are not requested at this time.

Recommendations

The following documentary components should be prepared:

1) A core or abstract specification of validation, which should cover all three S-100 data formats (ISO 8211, GML, and HDF5). This document should describe check classfication, syntax, operators (comparison, logical, and spatial), and general flow of check processing. This should be a document which can be referenced by individual product specifications, so that individual product specifications need not duplicate basic definitions like the spatial operations.
2) Generalized checks for each format. The ISO 8211 set has been developed (S100WG4 4.5) and is out of the scope of this paper; the GML and HDF5 sets are to be developed.

The nature of the above documents remains to be determined, including whether they should become part of S-100 or a separate specification. Spreadsheets may be the best vehicles for specifying the tests (especially the HDF5 set), but would be difficult to integrate into a text document.

Actions Requested

The S-100 WG is invited to:

1) Endorse the development of generalized validation checks for the S-100 GML and HDF5 formats on the lines described in this paper and the accompanying Annexes.
2) Subject to item (1), advise on possible variations from the outlines which may be needed for other product specifications.
3) Take additional action as appropriate.