

Paper for Consideration by S-100WG7

S-1XX Dataset (and Support File) Naming Convention

Submitted by:	IC-ENC
Executive Summary:	A consistent approach is required for the naming convention for S-1XX datasets (and support files) to ensure compatibility across different software and systems.
Related Documents:	S-100 5.0.0 Part 17, S-97 1.1.0, S-1XX product specifications: S-101 1.0.0, S-102 2.1.0, S-104 1.0.0, S-111 1.0.0, S-122 1.0.0, S-123 1.0.0, S-127 1.0.0, S-128 1.0.0 and S-129 1.0.0.
Related Projects:	n/a

Introduction / Background

IC-ENC is developing S-1XX services to support its members transition to S-100. The different approaches to the naming convention for S-1XX datasets (and support files, where applicable) has caused issues with the implementation of the ingest and registration process for multiple S-1XX products within the IC-ENC workflow tool.

Analysis/Discussion

IC-ENC has developed S-1XX Registration Checks to support the ingest and registration process for S-1XX products. A number of these checks focus on the dataset file name (and support file name, where applicable), described in S-100 Edition 5.0.0, Part 17-4.3 as:

Dataset naming shall follow a standard pattern to give implementers the assurance of unique names for incoming datasets.

XXXXXXXXXX.[EXT]

- *XXX is the product code (for example, 123 for Maritime Radio Services; 101 for ENC)*
- *YYYY is the producer code according to the Producer Code Register*
- *XXXX is an arbitrary length unique code in alphanumeric characters including any differentiating characters as required. The code shall be unique for the data producer (that is, different data producers may use the same code) and not re-used.*
- *EXT is the file encoding specific file extension*

Supporting resources shall follow the same naming convention.

Although the first seven characters in the dataset name are standardised, the eighth character onwards is open to interpretation regarding:

- number of characters,
- upper and lower case letters
- differentiating characters, e.g., (!"£\$%^&*@_), etc.
- extension shown as upper and lower case, e.g., .XML, .TXT, .h5

The S-1XX product specifications have implemented S-1XX dataset naming in different ways, albeit recognising that none are presently aligned to S-100 Edition 5.0.0. However, the current dataset (and support) file naming conventions allow for inconsistent file naming across the S-1XX products, as shown in Annex A based on current product specifications.

There is currently no guidance on the harmonisation of naming convention for S-1XX, and it is unknown whether no upper limit for the number of characters and the different characters used could seriously impact some systems.

Conclusions

It is noted that a core aim of S-97 is to assist in the creation of harmonised product specifications that are used within the e-Navigation eco-system. S-97 states that dataset naming should follow a standard pattern to give implementers greater predictability of incoming datasets.

The standardisation of file naming convention for S-1XX dataset names (and support files, where applicable) will improve machine readability and ease implementation. Harmonisation of the S-1XX naming convention will provide harmonisation across the product specifications, particularly between S-57 and S-101, and other S-1XX datasets.

Recommendations

A consistent approach to file naming convention across S-1XX product types is recommended for operational data. This includes:

- Harmonisation of dataset (and support file) naming for the 8th – *n*th characters.
 - Alphanumeric characters: upper case only
 - Maximum number of characters: tbd
 - Allowable differentiating characters: _ (underscore only)
- Standardised extension (EXT)
 - Upper case only (as S-101)
- Impacts/changes required:
 - S-100, Part 17-4.3 to provide clearer guidance for implementation across S-1XX product specifications.
 - S-97, A-6.2.18 Dataset naming rules
 - S-1XX product specifications to include harmonised dataset (and support file) naming convention.

The recommended approach above follows the S-101 implementation of dataset (and support file) naming convention, with .EEE as .H5 for S-102, S-104, and S-111, and .GML for S-122, S-123, S-127, S-128, and S-129.

101CCCC0000000000.EEE

The file name forms a unique S-101 identifier where:

- 101 - the first 3 characters identify the dataset as an S-101 dataset (mandatory).
- CCCC - the fourth to seventh characters identify the producer code of the issuing agency (mandatory for S-101). Where the producer code is derived from a 2 or 3 character format (for instance when converting S-57 ENC's), the missing characters of the producer code must be populated with zeros ("00" or "0" respectively) for the sixth and seventh characters of the dataset file name, as required.
- 0000000000 - the eighth to the maximum seventeenth characters are optional and may be used in any way by the producer to provide the unique file name. The following characters are allowed in the dataset name: A to Z, 0 to 9 and the special character _ (underscore).
- .EEE – new datasets and new editions use 000, updates start at 001 and increment until a limit of 999 (mandatory). Re-issues use the same number as the last Update applied to the dataset. Cancellations use the next sequential number from the previous Update applied to the dataset

The benefits of a fully harmonised naming convention for S-1XX products will ensure standardised system implementation of S-1XX product types and compatibility across different software and systems.

Action Required of S-100WG7

The S-100WG is invited to:

- a. Discuss the recommendations of this paper to agree a harmonised approach to S-1XX naming convention across the S-1XX product specifications.

Annex A - S-1XX Dataset file naming inconsistencies

		S-10x Dataset file naming inconsistencies																							
PS	Format	Char 1	Char 2	Char 3	Char 4	Char 5	Char 6	Char 7	Char 8	Char 9	Char 10	Char 11	Char 12	Char 13	Char 14	Char 15	Char 16	Char 17	Char 18	Char 19	.E	E	E	Comments	
S-101	101CCCCØØØØØØØØØØ.EEE	1	0	1	C	C	C	C	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø			0	0	0	a	
S-102	102PPPPØØØØØØØØØØ.H5	1	0	2	C	C	C	C	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	.H	5		b	
S-104	104CCCC "unrestricted characters" .h5 or .hdf5	1	0	4	C	C	C	C	unrestricted characters													.h	5		c
S-111	111CC "unrestricted characters" .h5 or .hdf5	1	1	1	C	C	unrestricted characters													.h	5		d		
S-122	CCNPI122XXXXXXXXX.GML	C	C	N	P	I	1	2	2	X	X	X	X	X	X	X	X				.G	M	L	b	
S-123	CCNPI123XXXXXXXXX.GML	C	C	N	P	I	1	2	3	X	X	X	X	X	X	X	X				.G	M	L	b	
S-127	127CCCCXXXXXXXXXX.GML	1	2	7	C	C	C	C	X	X	X	X	X	X	X	X	X	X			.G	M	L	b	
S-128	128CCCCXXXXXXXXXX.GML	1	2	8	C	C	C	C	X	X	X	X	X	X	X	X	X	X			.G	M	L	b	
S-129	129XXXXYYYYYYYY.GML or .gml	1	2	9	X	X	X	X	Y	Y	Y	Y	Y	Y	Y	Y					.G	M	L	e	
Key:		dataset, i.e. 101																							
		producer code																							
		mandatory character																							
		optional to be used in any way by the producer to provide unique file name																							
		NPI - nautical publication information																							
		extension																							
		not used																							
Comments:	a	.000 denotes NC/NE and 001 - 999 denotes Updates (A to Z, 0 to 9, and _)																							
	b	(A to Z, 0 to 9, and _)																							
	c	or .HDF5, or .hdf5 - lower case allowable in PS																							
	d	or .hdf5 - lower case allowable in PS																							
	e	or .gml (A to Z, 0 to 9, and _)																							