**TSMAD 25-4.6.3**

**Paper for Consideration by TSMAD**

**Auxiliary Files and S-101**

| | |
|---|---|
| ***Submitted by:*** | UK |
| ***Executive Summary:*** | Proposal for encoding auxiliary file information (TXT and TIF) in S-101. |
| ***Related Documents:*** | |
| ***Related Projects:*** | |

**Introduction / Background**

This paper proposes a change in the way text and picture files are deployed in S-101. A number of reasons are given for this proposal, mainly concerned with the ease of installation and management by end user systems and the complexity of maintaining external files within datasets.

## Analysis/Discussion

1. Auxiliary files are used in current S57 ENC datasets to hold text and picture information in ASCII and .tif formats. They are encoded included within exchange sets as "external files" where the name of the file is held in a TXTDSC/PICREP attribute value.

2. Currently, across the global ENC datasets, some 24,000 text and 300 picture files are encoded within cells. When the same text file is required by different cells copies are made in the exchange set media (although it would be possible to hold TXT/TIF files centrally this is rarely done in practice). There are ~11,000 unique text files within the global dataset with an average size of 900 bytes. Of the 24,000 total text files approx a quarter of them are below 300 bytes. The total file size of all the TXT files in the global dataset is approx 24Mb, or <1% of the total size of the entire ENC dataset. When the same text file is required by the same cell it can simply be referenced again

3. Each of the the 24,000 files are referenced on average only 3 times in each cell. This means there is repetition within a cell for the information. In practice this tends to be overuse of CTNAREs with external files, an acknowledged issue with current ENC encoding. Picture files are very rarely referred to more than once.

4. There are a number of complications and shortcomings of the current model, namely:

    a. A naming scheme, updating methodology, installation and management scheme is defined in parallel with the existing feature/attribute methodology. TXT/TIF files need to be updated and managed using a different scheme from that which is implemented for other S57 feature objects. Moreover a naming scheme for the files themselves has to implemented, validated and managed even though the file names themselves hold no useful information. Indeed NTXTDS attributes are even encoded in unicode as well as the file contents themselves.

    b. There is no top level map in an exchange sets metadata to denote which external file belongs with which cell file. This information is not held in the current catalog.031 file nor in the cell's metadata.

    c. A whole set of validation rules are required, therefore, that files have correct content, that no uniqueness conditions are broken (same filename / different content) and have to be managed by producing countries, validating bodies and implementing systems.

5. These shortcomings have resulted in end user systems where there is a wide variety of behaviour for

management of external files, where the end user is confronted with TXT/TIF filenames rather than the information within them and where there is a substantial risk of missing important information.

6. In summary – S57s design includes external files which then require a parallel system of maintenance for all entities in the data chain and there is little benefit for the end user in this data being held externally to the main dataset.

**Conclusions**

7. This paper proposes that within S-101 the content of the TXT and TIF files are embedded directly into the dataset file. Within iso8211 it is possible to encode long ASCII characters and binary data directly. The content of the files would then just be attribute values rather than embedded within an external file requiring its own change management system. The existing methods of feature attribution, update and management then take over and no other method of control or management is required. Iso8211 is perfectly able to cope with this data directly without requiring an outdated mechanism of embedding within a file. End user systems then have certainty over data contents and a much easier method for inport, management and display to the end user.

8. The issue of repetition within dataset files should be fixed by use of information types. Where a piece of text is required by more than one feature object it should be coded by reference to an information type within the dataset.

9. This paper also proposes that the S-101 model could be adopted at the S-100 level meaning that all "extra" dataset content could be encoded within iso8211 for other S-10x data products. Consideration should be given to reducing the number of other formats in the interests of simplicity for future data products.

10. Within other encodings (such as GML) it may well be necessary to continue to use other external files to encode image data, e.g where binary formats are required.

**Recommendations**

a) TSMAD to note the contents of the paper and to draft/propose amendments to the iso8211 encoding for S-101 and S-100 aimed at representing the current TXT and TIF contents directly. This would use iso8211's method for representing ASCII text, Unicode text and arbitrary length binary data.

**Justification and Impacts**

1. Reduction in effort for data producers, tool manufacturers, data validation, distribution/
2. Reduction in confusion for end user.

**Action Required of TSMAD**

The TSMAD is invited to:

a. endorse the recommendation to remove external files from S-101 and to investigate whether any external files are necessary for other S-10x products