Paper for consideration by TSMAD

Harmonization of text information model

Submitted by:	SNPWG / NIPWG
Executive Summary:	This paper describes a proposal for harmonization of the model for general text information.
Related Documents:	(1) S-101 DCEG (Baseline version – April 2014); (2) Accompanying detailed proposal TSMAD29/DIPWG7-10.5B
Related Projects:	(1) S-101; S-122; S-123; other product specifications for nautical information.

1   Introduction/Background

The development of the MPA and Radio Signal Product Specifications and the initial check of the S-101 DCEG baseline version identified the need to simplify the SNPWG data model for textual information. While the two attributes for encoding general text information in the S-101 model (**information** and **textual description**) might be sufficient to encode text information which is present on paper charts, nautical publications datasets often need more expressive modelling of text. On the other hand, allowing three different ways to model text may be unnecessarily complicated. The simplified approach would employ a uniform method for all kinds of textual information, while allowing an abbreviated version of the model in certain cases. This paper describes the proposed common model.

2   References

ISO 19115-1:2014: Geographic Information – Metadata – Part 1: Fundamentals, 2014-04-01.

3   Discussion

3.1   Summary of revised model

The current model of general text information is revised as follows:

- Complex attributes **information** and **textual description** are merged into a single complex attribute (which retains the name **information**).

- The new **information** attribute also binds the following new sub-attributes:

  - An attribute for indicating location of a specific paragraph, clause, etc., within a larger text file.

  - An attribute allowing (optional) encoding of a headline for the content, e.g., a section title.

- A new complex attribute **text content** is introduced which has as its sub-attributes the new **information** attribute plus other optional sub-attributes which describe the content and provide links to online resources.

  - Attribute **category of text** is intended for indicating the level of detail.

  - Attribute **online resource** is almost identical to ISO 19115-1:2014 class CI_OnlineResourceInformation (the difference is that the ISO 19115-1 sub-attribute *name* was changed to *nameOfResource*, since S-101 already defines *name*).

  - Attribute **source indication** indicates the source of the text content. **Source indication** and its sub-attributes are common to this proposal and the data quality harmonization proposal.

The revised model is shown in the figure below. Detailed specifications are provided in an accompanying draft proposal for revisions to S-101 feature and attribute types.
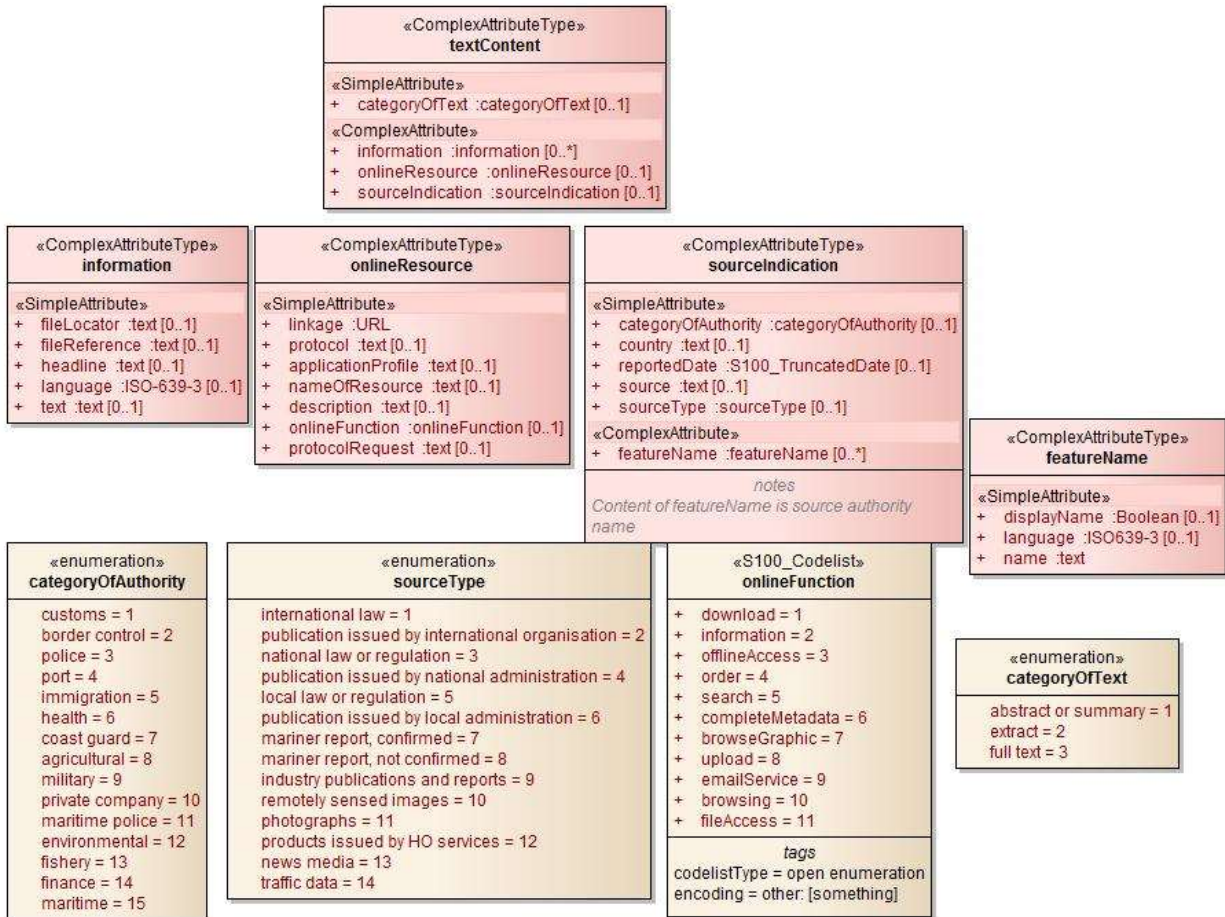
«ComplexAttributeType»
textContent

«SimpleAttribute»
+ categoryOfText :categoryOfText [0..1]

«ComplexAttribute»
+ information :information [0..*]
+ onlineResource :onlineResource [0..1]
+ sourceIndication :sourceIndication [0..1]

«ComplexAttributeType»
information

«SimpleAttribute»
+ fileLocator :text [0..1]
+ fileReference :text [0..1]
+ headline :text [0..1]
+ language :ISO-639-3 [0..1]
+ text :text [0..1]

«ComplexAttributeType»
onlineResource

«SimpleAttribute»
+ linkage :URL
+ protocol :text [0..1]
+ applicationProfile :text [0..1]
+ nameOfResource :text [0..1]
+ description :text [0..1]
+ onlineFunction :onlineFunction [0..1]
+ protocolRequest :text [0..1]

«ComplexAttributeType»
sourceIndication

«SimpleAttribute»
+ categoryOfAuthority :categoryOfAuthority [0..1]
+ country :text [0..1]
+ reportedDate :S100_TruncatedDate [0..1]
+ source :text [0..1]
+ sourceType :sourceType [0..1]

«ComplexAttribute»
+ featureName :featureName [0..*]

notes
Content of featureName is source authority name

«ComplexAttributeType»
featureName

«SimpleAttribute»
+ displayName :Boolean [0..1]
+ language :ISO639-3 [0..1]
+ name :text

«enumeration»
categoryOfAuthority

customs = 1
border control = 2
police = 3
port = 4
immigration = 5
health = 6
coast guard = 7
agricultural = 8
military = 9
private company = 10
maritime police = 11
environmental = 12
fishery = 13
finance = 14
maritime = 15

«enumeration»
sourceType

international law = 1
publication issued by international organisation = 2
national law or regulation = 3
publication issued by national administration = 4
local law or regulation = 5
publication issued by local administration = 6
mariner report, confirmed = 7
mariner report, not confirmed = 8
industry publications and reports = 9
remotely sensed images = 10
photographs = 11
products issued by HO services = 12
news media = 13
traffic data = 14

«S100_Codelist»
onlineFunction

+ download = 1
+ information = 2
+ offlineAccess = 3
+ order = 4
+ search = 5
+ completeMetadata = 6
+ browseGraphic = 7
+ upload = 8
+ emailService = 9
+ browsing = 10
+ fileAccess = 11

tags
codelistType = open enumeration
encoding = other: [something]

«enumeration»
categoryOfText

abstract or summary = 1
extract = 2
full text = 3

Figure 1. Revised model of text content

3.2   Justifications for revisions

- A common model of text information allows reuse of code modules and standardization of user interfaces for all data products using the model.

- Merging **information** and **textual description** means application schemas become a little simpler since they can bind one complex attribute for general text information instead of two, but the expressive power of application schemas is unchanged.

- Adding the capability to indicating location within a larger text file allows cartographers to reduce the number of support files since it becomes possible to identify a specific chunk of text within a support file in a structured format like HTML. Sometimes source material is already available in structured text format and cartographers may be able to use such files without further editing.

- Adding the abilities to encode headlines and categorize content allows more user-friendly display of some forms of text information, e.g., sections from national shipping regulations. Some hydrographic offices feel obliged to include the full text of a regulation because it is difficult or impossible to encode all the variations in language and conditions. Others desire to encode summaries or brief extracts. The attribute **category of text** allows producers to designate the level of detail of the encoded text, so that multiple levels of detail can be made available in the same dataset.

- Adding **online resource** allows indication of a web address where a document is made available e.g., by a local government, and also provides for programmable queries to web services, for future product specifications and applications that need such.

3.3 Use of model

The figure below provides examples from the S-122 (Marine Protected Areas) and S-123 (Radio Services) application schema.
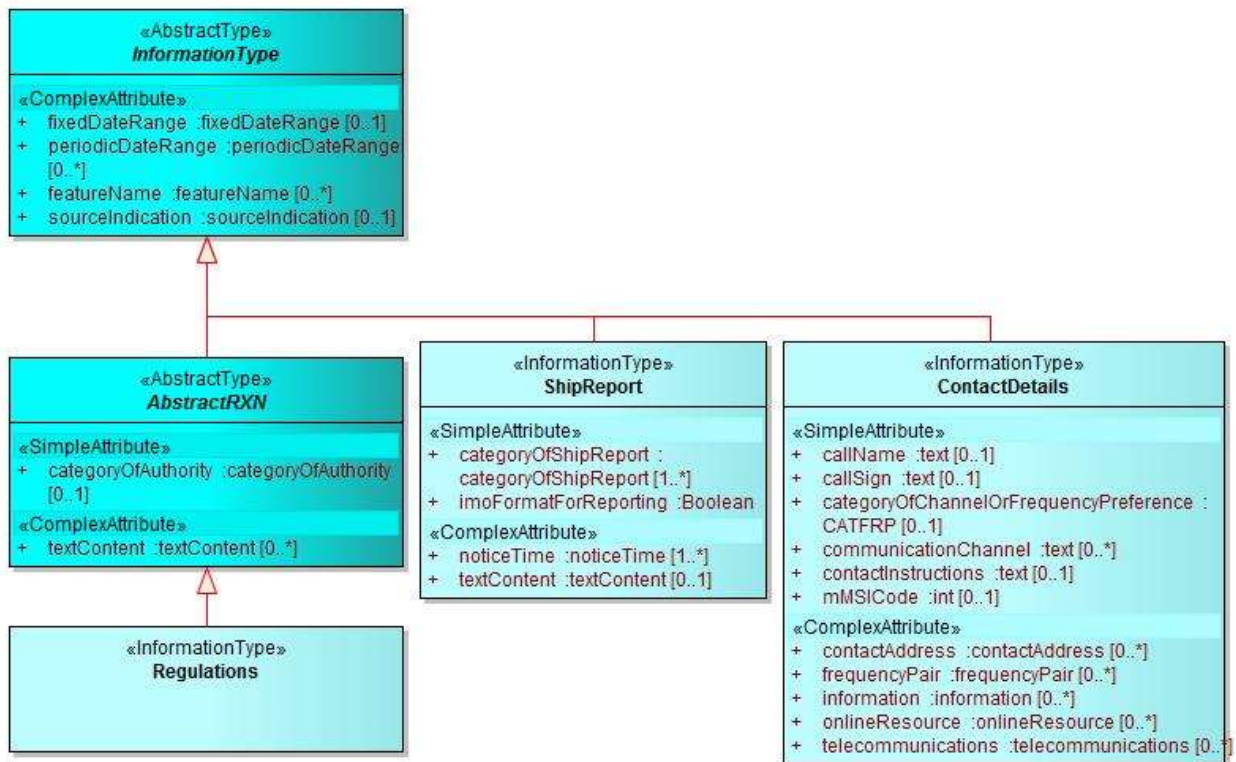

Figure 2. Examples of use in application schema

Example 1

Information type Regulations is intended for encoding regulations such as extracts from shipping regulations. It inherits the complex attribute **textContent** from its parent. This allows regulations to be encoded either:

- As embedded text in Regulations -> textContent -> information -> text.

- In an HTML support file. The file is named in Regulations -> textContent -> information -> fileReference.

- As the URL of the most recent online version of national shipping regulations, using Regulations -> textContent -> onlineResource -> linkage for the web address.

Example 2

Information type ShipReport describes the type, reporting times, and format of a ship report. The precise format and required components are described in a support file named by ShipReport -> textContent -> information -> fileReference.

Example 3

If certain conditions are satisfied it is possible to abbreviate the model even more. This is the case if the text information is a short string that effectively acts as a substitute for an enumerated attribute or "listed value" that is not part of the model. In these situations the other sub-attributes of "**textContent**" (e.g. **categoryOfText**, **onlineResource**, and **sourceIndication**) will not be used. The full "**textContent**" model is superfluous in these cases and adds unnecessary overhead and the complex attribute **information** can be bound instead.

For example in Figure 2, the feature **Contact Details** does not bind "**text content**". Instead it binds the attribute "**information**" which can be regarded as acting as a proxy for a hypothetical "category of contact" or "description of contact" attribute (neither of which is defined). For **Contact Details**, the additional text information is encoded as:

```
<ContactDetails>
  <information>
    <text>Supplementary information for Jussland MRCC RT (MF) radio</text>
  </information>
  <frequencyPair>
    <frequencyShoreStationTransmits>2182000</frequencyShoreStationTransmits>
    <frequencyShoreStationReceives>2182000</frequencyShoreStationReceives>
    <frequencyShoreStationTransmits>8291000</frequencyShoreStationTransmits>
    <frequencyShoreStationReceives>8291000</frequencyShoreStationReceives>
  </frequencyPair>
</ContactDetails>
```

3.3.1   Source indication

This complex attribute is also part of the NPUB data quality model, and is described in that proposal. Its role in the text information model is to indicate the source of the text information.

3.3.2   Suggested application of new model to S-101

The April 2014 baseline S-101 DCEG defined **information** and **textual description** as separate (complex) attributes both of which are bound only to Supplementary Information. The S-101 model can be updated in one of the following ways:

- Replace the bindings of **information** and **textual description** to **Supplementary Information** by the single complex attribute **text content**; or,

- Replace the bindings of **information** and **textual description** to **Supplementary Information** by the (revised) complex attribute **information**, if TSMAD believe that ENCs will never need **category of text**, **online resource**, and **source indication**.

Both are minor changes, the second in particular (binding of new attribute **information**) is no more than a merger of two existing complex attributes. (The sub-attributes **file locator** and **headline** have multiplicity lower bound 0, and their use can be restricted.)

4   Conclusion

The proposed changes to the text information model results in a more generalized common model that can be used across multiple IHO product specifications including nautical publications datasets as well as allowing capture of additional details about text. A common model of text information allows reuse of code modules and standardization of user interfaces for all data products using the model.

5   Actions Requested

TSMAD is invited to:

- review the accompanying detailed proposal for modeling text information and suggest amendments as appropriate;

- interact with SNPWG/NIPWG to harmonize the modeling of text information in nautical publications and S-101 application schemas to the extent possible