WORLD METEOROLOGICAL ORGANIZATION   IPET-DRMM-I / Doc. 3.3 (1)

COMMISSION FOR BASIC SYSTEMS   (18. 6. 2013)

-----------------------------   ------------------------

FIRST MEETING OF

INTER-PROGRAMME EXPERT TEAM ON   ITEM 3.3
DATA REPRESENTATION MAINTENANCE AND MONITORING

                                                      ENGLISH ONLY
TOKYO, JAPAN, 1 - 5 JULY 2013

Toward BUFR Edition 5: a discussion document

Submitted by

Yves Pelletier (Canada),

_____

Summary and Purpose of Document

This document provides a discussion on the status of BUFR and a table of
proposed characteristics of an eventual BUFR Edition 5.

_____

ACTION PROPOSED

The meeting is requested to note the information, discuss and clarify the proposals as required, add
any new proposals and discuss next steps.

Introduction

This document is produced at the request of the Inter-Program Expert Team on Data Representation and codes, mainly following discussions in Melbourne (2011) and Exeter (2012). It is not intended as a statement of requirements. Rather, it aims to foster awareness of the context and motivation for a new BUFR edition, and to provide material for a focused discussion of the eventual feature set of BUFR Edition 5. It is hoped that the actual list of requirements will derive from these discussions. Therefore, for the purposes of this document, we will refer to requirement candidates or proposals and not to requirements as such.

Approach

We clustered requirement candidates around certain themes. Metadata interoperability is prominently considered to be a theme upon itself. Important work was performed by the Expert Team on Metadata and Data Interoperability on this topic in relation to BUFR, and a number of proposals stem directly from that work. The other themes are more general, respectively grouping functional and procedural enhancements. It is not certain that all requirement candidates will make the final list but some have been noted as being necessary in order to meet high-priority objectives of the WMO in relation to metadata interoperability.

As discussed at the IPET-DRC meeting in Exeter (2012), we made a deliberate effort to differentiate explicitly between a requirement, which should be expressed in general terms, and a proposed implementation of a requirement, which is in fact a proposed solution to meet a perceived requirement.

Changes in edition number are the recognized mechanism whereby enhancements to BUFR that require modifications to encoding/decoding software implementations are introduced into the code form specification. This is not undertaken lightly, as the implementation of software changes in operational systems involves significant resources in each Processing Centre. In this instance, some of the proposed requirements related to metadata interoperability would lead to changes more far-reaching than have ever taken place so far in previous BUFR Editions. This will be discussed in more details in the table in annex.

Fundamental BUFR requirements

BUFR was designed with certain characteristics in order to meet requirements that are imposed by the nature and constraints of meteorological data, the meteorological enterprise and the demands of worldwide meteorological data processing. Proposed requirements should preserve or enhance the way BUFR can achieve these characteristics, but there is also room to consider these fundamental characteristics while taking into account the evolution of the operating environment in the 25 years or so since BUFR was first devised.

The defining characteristics of BUFR are:

1. Universality in the representation of data and metadata (achieved through the BUFR data representation syntax and the various code tables)
2. Extensibility (meaning that new data elements or sequences can be added without changes to the encoding/decoding software; achieved by Table-driven approach)
3. Portability across computer architectures (achieved by use of unsigned octets, scale,

reference value and bit-width)
4. Compactness (achieved by Table-Driven approach and representation of code values in binary form, using as few bits as possible).
5. Lossless conveyance of information (no lossy compression is used. Also, while current editions do not focus on data integrity in the technical sense of the word, there is a concern for sufficient precision to preserve the accuracy of the data being conveyed.)
6. Focus on operational data exchange between data processing centres

And in light of discussions at the joint ET-MDI and IPET-DRC meeting in Exeter, we might add a new, proposed fundamental requirement:

7. *Interoperability with ISO standards for data and metadata representation.*

BUFR edition 5 – Requirement candidates and discussion

	Theme	Proposed Requirement	Proposed implementation	Discussion/Commentary
1.0	**Metadata interoperability**			
1.1		Logical Data Model	The development of a Logical Data Model (Application Schema) should be the initial activity in developing a new data exchange requirement. Endorsement of this principle should be reflected in changes to the current BUFR regulations. (Proposer: UK Met Office)	The Joint IPET-MDI & IPET-DRC workshop held at the UKMO headquarters in May 2012, agreed to make the following recommendation to OPAG-ISS. *IPET-MDI & IPET-DRC recommend in principle the approach of using a Logical Data Model to drive the development of new data representations. They ask that OPAG-ISS recommend to CBS that the new approach be considered and endorsed. If the proposal is found to be acceptable , CBS should task an expert team to further develop the approach by 2016.*
1.2		Enforced use of templates in product development	The use of Table D sequences in the representation of new data types , should be regulated and enforced through changes to the current BUFR regulations. (Proposer: UK Met Office)	**UK Met Office**: In making the above recommendation (1.1) it was noted that there is currently no regulation to enforce the use of Table D sequences. The resultant lack of structure has allowed publishers to introduce a degree of variation in their encoding . **Canada**: Implementation of this proposal would incur a loss in flexibility in coding sequences for data representation. As in programming languages, we are a little leery of attempting to enforce good programming through regulations of the code form itself. In this particular case, might it not be a case of needing a QA procedure to ensure the fitness of descriptors sequences for a particular purpose?
1.3		Consistent approach to handling of generalized coordinates and Polymorphism	The regulations for BUFR edition 5 must stipulate that all table D sequences used must be opened and closed by the occurrences of the same Generalised Coordinate and / or Significance Qualifier. This requirement supersedes regulations (94.5.3.3 - 94.5.3.6) for edition 4. (Proposer: UK Met Office)	**UK Met Office**: Under current BUFR regulations (94.5.3.3 - 94.5.3.6) element descriptors remain in effect until either superseded by redefinition or by an re-occurrence of the opening descriptor which has been set to "missing" (all bits set to "1"). The work carried out in the development of a "sympathetic" Logical Model for the OPMET data has shown that this practice will have unanticipated side effects. It has been accepted that a majority of UML classes will be created or adapted from existing Table D sequences, and for these to be effective they must be opened and closed by occurrences of the same General Descriptor / Significance qualifier. **Canada**: This is a desirable feature to improve machine-readability of BUFR as well as inter-operability. However it should be noted that such a regulation would invalidate many current Table D descriptors. Up to now, each new BUFR Edition preserved the format and previous contents of the Tables. This proposed requirement would change this, requiring a refactoring of many (most) existing Table D sequences. The validity of some popular Table D sequences containing only coordinate descriptors, such as 301024 (lat/lon/elevation), would also come into question. For each such sequence, a cancelling mechanism would become necessary, perhaps through a Table C operator or through new regulations governing a new type of coordinate class Table D descriptor. Also of interest, this excerpt of an e-mail to the Team from **Atsushi Shimazaki of the WMO Secretariat**: *as regards cancellation, I think it would be better to note that there are three types of coordinate descriptors (excluding Table C descriptors);* *1. Applicable to descriptors in general* *This type can define coordinates for a set of descriptors. I suppose majority of descriptors in coordinate classes are this type. This type should be cancelled implicitly or explicitly as mentioned above 2nd paragraph.*

				2. Applicable to a specific element descriptor In Class 02 (and possibly in other classes), there are descriptors that are connected to specific descriptors, such as 0 02 002. *Regarding this kind of descriptors, it is not clear whether they need cancellation EXPLICITLY, because they are not effective to descriptors other than relevant "specific descriptors", although Eva strongly disagrees with cancellation of Class 02 descriptors. If those do not need to be cancelled explicitly, we need to identify descriptors to be included in this type.* *Meaning of "contradiction" could also be discussed in this context. I agree with Milan's definition that redefinition and contradiction refers to the coordinate elements with same reference numbers. This is very simple.* *3. Applicable to "following" descriptor* *There are some descriptors with the word, "following" in names, such as 0 04 080 and 0 08 019,and similar descriptors are also seen in the Class 33. Quality information defined by the descriptors in Class 33 seems to act as a coordinate in the sense that they define "attributes" to the following data. As the item 2 above, it could be discussed whether those descriptors should be cancelled EXPLICITLY, because they define a coordinate applicable to only next descriptor.*
1.4		Short and long names for Table B & D entries	Two new columns will be required in both BUFR Table D and BUFR Table B providing full and abbreviated element names.	**UK Met Office**:The introduction of short & long name entries in Table D and B will prove to be beneficial in the creation of UML data models. Whilst shorter names are useful in UML, they may cause confusion to non experts. To counter these more expansive definitions will help non experts in understanding the terms being used.
1.5		Align coordinate descriptor behaviour and syntax with ISO terminology and functionality	Likely to require re-definition of coordinate descriptor classes	**Canada (from yp personal meeting notes)**: I highlighted this as meaningful in my meeting notes but the specifics might better be explained by data modelling experts. I jotted down the following, which may help understand what this is about: Modifier behaviour • Property • Property ID • Lat/lon • Position parameters • Vertical and temporal • Property role • Inheritance-discrimination If none of this makes sense we can just ignore it…
2.0	**Functional enhancements**			
2.1		Representation of values over extremely large dynamic range	Introduce an IEEE floating point representation operator in Table C.	**Canada**: one use-case would be concentration values for suspended or gaseous chemicals in the atmosphere. It is felt that IEEE representation is a well-known and applied standard which does not detract from the interoperability requirement for BUFR. A proposal by Canada was implemented experimentally in software by ECMWF and Canada, including compression.
2.2		Explicit separation of data types from units	Add a column to table B so that data type (numerical, code table, etc) is stated separately from units	**Canada:** semantically this makes more sense. I surmise the mashing up of type and unit was justified by the compactness requirement, but is that still the case?

2.3		Multilingual support for character strings	Support UTF-8 (proposer: Canada)	
2.4		Delayed sequence definition		**From yp personal meeting notes; proposer may be Gil Ross.**
2.5		Table D constants		**From personal meeting notes. What Table D constants are and how they would be useful is to be clarified if possible. (Gil Ross?)**
2.6		Increase name-space size for descriptor classes and descriptor numbers (XX and YYY in F-XX-YYY)	To be determined	**Canada:** In the current Table B, one notes that certain classes are filling up. This will soon be an issue when new parameters need to be defined.
2.7		Random or quasi-random access to data inside a message	To be determined	**EUMETSAT:** It would be useful (particularly when dealing with large compressed data sets) to know where the information associated with a given parameter starts, without having to access the data in a serial manner to find it. I mean that if, in the equivalent of the current Section 3, it was possible to say for any element where they can be found in the message. That way I could decode latitude, longitude and temperature without needing to wade through all the parameters in between. I could imagine something along the lines of a few bytes per element giving the offset into the BUFR data to the start of that element. For a typical data set we use about 100 elements so I would add about 400 bytes if all elements had this kind of info (it could be optional) – the benefit would be quasi random access to the parameters and the associated speed up in processing.
	Procedural enhancements			
3.1		Versioning and life-cycle of Table B and D entries	To Be Determined	**Canada**: While the current editions of BUFR allow for versioning of tables B and D, this was never enforced. Furthermore, assumptions were made about backward compatibility of Tables B and D versions, which were never in the specifications. An enforceable versioning mechanism would help ensure that decoders can reliably decode BUFR that was encoded with one version of the tables or another. On the question of life-cycle, an examination of Table B and D leads us to observe that there are many obsolete and, sometimes, flawed entries that are no longer used or usable. A mechanism for the removal of such items from the tables would help maintain the coherence and long-term viability of the tables.
3.2		A new methodology is needed in the new editions of BUFR and GRIB to ensure that decoders are always using the exact same table information as encoders.	To Be Determined	IPET-DRC report, item 8.5: *The current methodology of version numbers is inadequate because in many cases it is ignored or not fully implemented. Furthermore, most centres manage a local set of the WMO tables. Each of these local implementations may contain errors, and it is not currently possible to provide an objective guarantee of the tables' integrity at the time of encoding or decoding.* See also document submitted by US at Exeter meeting: http://www.wmo.int/pages/prog/www/ISS/Meetings/IPET-DRC_Exeter2012/Documents/IPETDRC-IV_Doc8-5_tableregistry.doc

3.3		Reference API	To Be Determined	Short of a reference implementation, it has been suggested a reference API might help with standardizing the implementation of BUFR.