Semantic Approach to WMO Codes

Enrico Fucile
European Centre for Medium-range Weather Forecasts
Shinfield Park, Reading, RG27HQ, U.K.

Keywords: BUFR, WMO Codes, Observations, IWXXM, METCE, EcCodes.

Abstract: Two initiatives are presented to make the WMO codes usable by a wider community: a new decoding library (ecCodes) and new WMO data formats based on a logical data model and ISO standards. Both are served by a web accessible registry as a source of semantics. The European Centre for Medium-range Weather Forecasts (ECMWF) has developed a decoding software (ecCodes) providing unified semantic access to the WMO codes through a web accessible registry. The software is very effective in providing easy access to the existing codes and to support conversion to a new set of WMO codes based on a logical data model (METCE) which enables wide interoperability support through a unified semantics conforming to ISO/TC 211. The first of such codes is IWXXM (ICAO Weather Information Exchange Model) which is the new worldwide standard for the exchange of meteorological information in the context of the Air Traffic Management.

1 INTRODUCTION

The World Meteorological Organization (WMO) is the specialized agency of the United Nations for meteorology (weather and climate), operational hydrology and related geophysical sciences. It has developed the Global Observing System (GOS) (WMO, 2010) as a coordinated system of methods and facilities for making meteorological and other environmental observations on a global scale. The observations made in GOS comprise in-situ measurements of atmospheric parameters from several platforms, satellite and other remote sensing observations. The data produced in these observations are exchanged on a co-ordinate global system of telecommunication facilities named GTS (Global Telecommunication System) (WMO, 2009).

GTS makes available to the National Meteorological Services and to their Numerical Prediction Centres a constant stream of observations on the state of the atmosphere used to produce weather forecasts at local and global scale. As an example the European Centre for Medium-range Weather Forecasts (ECMWF) is processing more than 300 million observational data elements every day to produce forecasts of weather and chemical composition of the atmosphere.

To make effective use of such a huge number of observations coming from a wide variety of sensors an internationally agreed data format designed for operational exchange is needed. At this purpose WMO has developed a set of code forms with a strong governance to make the exchange of information possible in a worldwide scale. This set of WMO codes constitutes an international standard for the exchange of information on weather, climate and hydrology. The importance of having a globally governed set of code forms for the exchange of observations is connected with the level of operational availability and quality of the data needed for the use of National Meteorological Service operational activities.

WMO codes are a complex set of coding standards developed in more than fifty years and in continuous evolution. They are divided in two distinct sets of alphanumeric and binary codes. The alphanumeric codes were developed for the transmission on telegraphic lines and are made to be written manually by a human observer and interpreted without machine support. With the evolution of telecommunications a new set of binary codes was developed to improve the quality and to provide compression algorithms which are needed in the transmission of the high volumes of data produced by satellites and some other remote sensing platforms as well as the

Fucile E..
Semantic Approach to WMO Codes.
DOI: 10.5220/0004843404090414
In *Proceedings of the 3rd International Conference on Sensor Networks* (MOEOD-2014), pages 409-414
ISBN: 978-989-758-001-7

forecast fields produced with numerical models. Moreover the progressive replacement of the manned observing stations with automatic stations has made the binary production of observation reports more convenient and effective.

The continuous evolution of the WMO codes and the heterogeneous nature of the observing system have resulted in a very complex set of codes which makes very difficult for a user to access the information requested in a consistent way across the various parts of the coding system. To improve consistency across the codes the Table Driven Code Forms (TDCF) (WMO, 2003) have been developed to replace the traditional alphanumeric codes (TAC).

2 TAC VS. TDCF

There are 76 different traditional alphanumeric codes (TAC) (WMO, 2012a). An example of TAC message is the Aerodrome routine meteorological report (METAR) (WMO, 2012a) which is the message of observation produced every hour by the airport observation station and used by the air traffic control to inform the pilots on the weather conditions at the airport. The code form is quite complex, but is made to be interpreted by a trained operator. Here follows an example:

```
METAR EDDF 120550Z 03015KT 1400
R07R/P2000N R07C/P2000N
R07L/1900U SN DRSN BR VV///
M04/M04 Q1000 R07L/11//90
R07C/15//90 R07R/15//90 BECMG
4000 NSW=
```

The code starts with the word METAR and ends with the sing "=". A set of numbers and letters are divided in groups which are recognized for their position and for their alphanumeric pattern. The element EDDF is the four character code for the airport, meaning that the observation refers to Frankfurt airport. The second element 120550Z says that the observation is for the 12th of the month at 5:50 UTC (the month and the year are not expressed in the message). The third element 03015KT means that the wind is coming from 30 degrees with a speed of 15 Knots. It is clear that decoding of each of the elements does not follow any general rule except the fact that information is split in groups with different meaning.

Another example of TAC is the SYNOP (Report of surface observation from a fixed land station). An example of this kind of message is the following:

```
AAXX 13094 03002 45462 /0514
10097 20073 30238 40256 58011
90850 333 88/11=
```

Where AAXX is the start of the report of this kind and "=" is the end. Only numbers and "/" are allowed in the body of the message, which are grouped in groups of 5 with exception of the groups marking new sections like the group 333 which denotes the start of section 3. The first element 13094 means that the observation is valid for the 13th of the month (no explicit indication of month and year) at 9 am and the number 4 means that the wind speed is observed with an anemometer and is reported in knots. Decoding the elements of the message is out of our scope we only want to point out that each group has a different meaning and each number within the groups has quite complicated decoding rules which are different group by group and code figure by code figure.

A coding system like the one used in the TAC without general rules, in which each element has different decoding rules is difficult to extend, to maintain and makes impossible the task of producing a general decoder. To overcome these limitations of the alphanumeric codes it was decided to produce a new system based on a unique set of tables providing a list of elements reusable in different contexts. A unique set of rules to decode a message were also provided in a form that is possible to implement decoding software to access information from the message. The fundamental concept of these new codes called Table Driven Code Forms (TDCF) was the separation between the coding rules and the elements used in the code forms. The rules are generic and apply to all the different types of messages, while the tables of elements and sequences are provided as external support to the decoding software in the form of a palette of elements to be reused in different contexts and a set of sequences of elements with a special meaning to be used in the definition of a single message. With the TDCF approach is possible to produce a decoder implementing the decoding rules, which can take as input the Tables and decode the message, allowing a big flexibility and extensibility connected with the fact that new messages can be defined through new elements and sequences without changing the decoding software.

WMO has developed two different TDCF named BUFR (Binary Universal Form for the Representation of meteorological data) and GRIB (General Regularly-distributed Information in Binary form). To limit the scope of this paper we will consider only BUFR which is also the more generic of the two formats.

BUFR (WMO, 2012b) was developed as a continuous bit-stream made of a sequence of octets (1 octet = 8 bits). A general structure of the message was defined and general rules for the decoding where based on tables of elements in which the number of bits used by that element and the parameters to get an integer or a floating point value from the content of the bits is given.

The message is composed of five sections. The first three sections can be considered headers and the last section is an end of message. The information is coded in section 3 and 4. Section 3 is a list of descriptors each of which occupies 2 octets and is made of three parts: F (2 bits), X (6 bits) and Y (8 bits). The sequence F-X-Y identifies uniquely the meaning and decoding rule for the descriptor. When F=0 the descriptor is a simple element and is coded as an integer I of N bits. To obtain the decoded real value I must be multiplied by 10^{-S} and a reference R must be added. A decoder will be able to decode the data by knowing the three coding parameters: S (scale), R (reference value), N (number of bits) for each descriptor with F=0. Tables of descriptors are provided in the WMO Manual on Codes [5] with the meaning and the coding parameters to be used by decoding software.

An example for the dew point temperature is given in table 1. When in section 3 of the BUFR message the element 0-12-003 is read by a decoder it means that in section 4 (data section) there are 12 bits in which an unsigned integer I is stored, that multiplied by 10^{-1} gives a “Dew point temperature” in Kelvin (the reference value is 0 in this example and therefore nothing needs to be added to the result).

BUFR decoding rules comprise also descriptors with F=1, 2 or 3. F=1 called replication descriptors and are used to realise repetitions of groups of descriptors. Descriptors with F=2 are operators acting on the following descriptors by changing their meaning or their decoding parameters. Descriptors with F=3 are called sequences and are used to define sequences of descriptors having a particular meaning and used in many cases to build other sequences. It is not the scope of this paper to give a detailed and comprehensive description of the decoding rules of BUFR. We only want to highlight that BUFR is based on general rules, which are very complex in some aspects and are based on external tables of descriptors which are published in the WMO Manual on Codes and in text or XML format from WMO web site www.wmo.int.

3 A SEMANTIC APPROACH TO BUFR

The approach implemented in BUFR provides a great flexibility in the definition of new messages and therefore in dealing with the growing set of sensors and observation types. However the connection between the BUFR code (0 12 003 in table 1) and its meaning has forced a generation of scientists to learn the code figures and to use them in their software. The fact that the meaning is connected with the code figure and that the code tables are external and can be easily replaced with convenient tables compiled by the user has made the use of BUFR code very difficult. The usability of a coding system as BUFR is very important. Indeed the fact that decoding is a process based on code numbers can prevent the use of useful information by the scientific community and slow down the implementation of new observations in the processing system of a numerical prediction centre. There are several initiatives to make BUFR more users friendly. They are based on the importance of giving clear semantics to the information in the message in a way that it will be easily accessed by any user and that the interoperability of different systems will be improved.

ECMWF is producing new decoding software (ecCodes) providing semantic elements to the user rather than numeric codes. The semantics used by ecCodes is going to be exposed in a web accessible registry to the benefit of a wide community and to make the decoding software independent from the semantic source which will be external and accessed by the software itself. On another initiative WMO has started an activity of review of the codes to enforce the semantics in the design of the coding systems using UML modelling to produce data models which are semantically rich and more

Table 1: BUFR table B entry for dew point temperature giving meaning and decoding parameters for the element

Code	Name	Unit	Scale	Reference value	Number of bits
0 12 003	Dew point temperature	K	1	0	12

suitable to work in an integrated web environment. The two initiatives aim to address the problem in two different time scales: ecCodes and the web registry are an attempt to give better accessibility to the existing codes, while the model based development of codes is a longer term initiative which will require the replacement of the current code forms with new data formats designed with different principles.

3.1 Decoding based on Semantics

The new decoding software developed at ECMWF, called ecCodes, is based on the concept of making a clear separation between the coding (binary) level and the semantic level of a BUFR message.

The software is written in C and provides bindings for FORTRAN, Python and Perl. It also provides a set of command line tools to perform simple operations on the messages.

ecCodes is not only a BUFR decoder. It is a general purpose decoder able to decode with the same function calls alphanumeric and binary WMO messages. The software is made of two main components: decoding engine and decoding rules. The engine is written in C and able to read the appropriate decoding rules from text files while parsing a binary or alphanumeric input message. The decoding rules are written in a language interpreted by the decoding engine and they are cached for efficiency reasons. Writing new decoding rules is fairly simple, but is not part of the work that a user has to do to decode a message. The decoding rules are provided with the library and are used at runtime while decoding messages. The interface provided to the user is based on a key/value approach. In all the available bindings there is a get function which is getting the value of an element of information associated to a key name. As an example to access the “dew point temperature” in a METAR or in a SYNOP or in a BUFR message, the function call is the same for the three types of messages:

```
codes_get(m,’dewPointTemperature’
,dewPointTemperature)
```

where `m` is the message previously loaded with an appropriate function, `dewPointTemperature` is the key name and the third argument is the variable to which the value found in the message is assigned. The example is of a FORTRAN function, but also Python, Perl and C equivalents are provided with similar syntax.

The function is actually looking for the element 0-12-003 in BUFR or for the corresponding elements in METAR and SYNOP to provide the value of the dew point temperature.

There are two problems that the software has to address in this operation. There can be more than one `dewPointTemperature` element in the message and the association with the unified semantics is a complex and difficult operation requiring a deep analysis of the BUFR tables and of the alphanumeric coding because they were not designed to support a unified semantics.

The first problem is solved with a specification of the information element. As an example if there are two different `dewPointTemperature` elements corresponding to different hours of the day it will be possible to distinguish them with

```
codes_get(m,’/hour=6/dewpointTemp
erature’,dt1)

codes_get(m,’/hour=9/dewpointTemp
erature’,dt2)
```

where `dt1` and `dt2` are the values of dew point temperature at 6 and 9 respectively.

An analysis of the code forms is being performed with the aim to provide a durable link between the numeric codes and the key names allowing the semantic access. It is evident that if the results of this analysis are made available only to the decoding software this will have the drawback of limiting the semantic access to ecCodes. With the aim of building a semantic of general use and accessible to any decoding software WMO is developing a web accessible registry to make the codes and its semantics publicly available and persistent. ecCodes will use the WMO codes registry as semantic source for the message decoding and will provide to the user a quick way of inspecting the meaning of each element of the message by navigating the registry. The WMO codes registry is already accessible at codes.wmo.int and provides at the moment only support for a small part of the WMO codes (the aviation codes) for which there was a requirement of building an XML format and supporting it with web accessible registry. The WMO codes registry will be soon developed to support BUFR.

3.2 Building Semantics in the Message Structure

It is sometimes very difficult to provide semantic access to a coded message when it was developed without the purpose of giving to it a clear semantic structure. Therefore WMO is planning to change the way a code form is designed by developing first a logical data model and deriving the corresponding physical coding with an automatic process. With this technique a logical model providing clear semantics is developed first and from this model a physical coding is derived by means of an automatic process. The same logical model can therefore produce BUFR and XML format implementing the same semantics. This will allow users of two very different coding systems like BUFR and XML to share the semantics and therefore to work independently from the coding format.

The process of deriving a data format from a logical data model has been applied by WMO for the first time in the messages used for aviation purposes: METAR (Aerodrome routine meteorological report)/ SPECI Aerodrome special meteorological report), TAF (Aerodrome forecast) and SIGMET (Significant Meteorological Information) [4]. The project is being developed by the WMO Task Team on Aviation XML (TT-AvXML), and the first release version of the logical model and the XML schemas, automatically derived from the model where publicly released in September 2013 [1].

A model has been developed METCE (Modele pour l'Echange des informations sur le Temps, le Climat et l'Eau) which is designed with the purpose of enabling semantic interoperability in the fields of weather, climate and water. At this aim it has been decided to make it compatible with the ISO/ TC 211 Geographic information/Geomantic. METCE makes use of the ISO 19156 Observation & Measurement to provide conceptual definitions of meteorological observations to be imported in domain specific application schemas.

An example of domain specific application schema is the ICAO (International Civil Aviation Organisation) Weather Information Exchange Model (IWXXM) which was developed by TT-AvXML to define the semantics for the aviation meteorological messages for which an XML version was requested by ICAO. [2]

To illustrate the advantage of using a unified semantics derived from a logical model to build the code format we take a single group from a METAR message reporting the wind observed at the aerodrome:

```
17006G12MPS
```

the meaning of this group is that the wind is blowing from the direction of 170 degrees with a mean speed of 6 metres per second and gusts of 12 metres per second. The corresponding XML conformal to the schema produced from the logical model is:

```
  <iwxxm:surfaceWind>
    <iwxxm:AerodromeSurfaceWindForecast variableWindDirection="false">
      <iwxxm:meanWindDirection
uom="deg">170</iwxxm:meanWindDirection>
      <iwxxm:meanWindSpeed
uom="m/s">6</iwxxm:meanWindSpeed>
      <iwxxm:windGustSpeed
uom="m/s">12</iwxxm:windGustSpeed>
    </iwxxm:AerodromeSurfaceWindForecast>
  </iwxxm:surfaceWind>
```

this construct is using terms which are defined in the logical model and are linked to a unified semantics published in the WMO codes registry at codes.wmo.int. The advantage of basing the code system on a unified semantics is that the terms are publicly accessible, reusable in different contexts and used with the same meaning in different coding systems. It is possible indeed to produce a BUFR format from the same logical model to allow the use of the same semantics from a very different coding standard.

4 CONCLUSIONS

The evolution of WMO codes used to exchange observations from a wide variety of sensors and platforms is a continuous process. The original alphanumeric codes where based on specific and fragmented decoding rules based on the meaning of sequences of numbers and letters with lack of a general unified semantics. TDCF were introduced to provide general decoding rules separated from the specific definition of the elements of information.

The large and growing number of different observation types requires a unified semantics to make use of the information in different contexts. There are several initiatives to provide access to the observations through a unified semantics. We have discussed two of these initiatives: development of a decoding software (ecCodes) exposing unified semantics for WMO TAC and TDCF messages, changing the development process of WMO codes making it dependant on the design of a logical data model providing a unified semantics. Both initia-

[1] Downloadable from http://www.wmo.int/pages/prog/www/WIS/wiswiki/tiki-index.php?page=AvXML-1

[2] The logical model can be visualised from http://wis.wmo.int/AvXML/AvXML-1.0/index.htm.

tives are based on a web accessible registry as a central source of semantic.

The two approaches complement each other because the decoding software will provide at a given extent the same semantics that will be built with a logical data model, provided that ecCodes will use the same terms that will be used in the development of the logical data model. At this aim a web accessible registry of semantic terms is being published by WMO under codes.wmo.int to provide the semantics for the decoding software and for the data modelling. This web registry will represent an authoritative source of terms maintained by WMO usable in different contexts and for different purposes.

WMO has started to develop a logical data model called METCE and the first specific domain application schema IWXXM for aviation meteorological data and the technique of designing a logical data model to automatically produce an XML schema has proven successful and will be applied to produce BUFR formats.

REFERENCES

Manual on Global Observing System, WMO No. 544 (2010)

Manual on the Global Telecommunication System, WMO No. 386 (2009)

Guide to WMO Table-Driven Code Forms FM 94 BUFR and FM 95 CREX, WMO 2003. Available from http://www.wmo.int/pages/prog/www/WMOCodes/Guides/BUFRCREXPreface_en.html

WMO Manual on Codes Part A - Alphanumeric Codes, WMO No. 306 Vol. I.1 (2012)

WMO Manual on Codes, WMO No. 306 Vol. I.2 (2012)