THE APPLICATION OF BINARY DATA REPRESENTATION IN METEOROLOGICAL SYSTEMS

by J. Böttcher
German Weather Service
Offenbach, Federal Republic of Germany

Summary: The amount of meteorological data is increasing, both from the observational side, due to new types of observational platforms like e. g. satellites, and the model output side, due to the use of models of higher complexity. The new binary codes GRIB and BUFR are designed to reduce the amount of data for exchange. However most of the present telecommunication network is not yet suited for the exchange of binary data. In principle data could be exchanged through a system of distributed data bases as well. Within a centre it might be necessary to transform telecommunicated data into a format which is more suitable for the internal use of data. However, this will depend on the objectives of the centre. Procedures to convert data from the data base formats to a user-friendly format are necessary in any case. The units of some elements as used in BUFR might call for a further conversion if data are represented on a plotted chart or a printed table.

1. INTRODUCTION

More detailed and complex numerical weather prediction models at present and in the future will need a growing amount of observational data. On the other hand the weather prediction models are producing more and more output data. According to the Second WMO Long-term Plan (1988-1997) the volume of observational data will be nearly doubled from 1985 to 1990. The Second WMO Long-term Plan contains also figures on the expected increase of the volume of data from one global forecast centre. In 1985 one global forecast centre provided 6 million octets per day. Estimates for 1990 and 2000 are 40 million octets and 200 million octets per day respectively. Therefore, direct computer-to-computer exchange formats like GRIB and BUFR will be used for the exchange and storage of large volumes of data both observational data and processed products between automated centres over high-speed channels, and a system of character-oriented codes (this may be the present WMO character codes as well as an improved version) will be used by centres not equipped for computer-to-computer exchange of data and products.

2. BINARY DATA-EXCHANGE FORMATS

Two binary codes are now defined in the WMO Manual on Codes:

FM 92 GRIB (gridded binary) - Processed data in the form of grid-point values expressed in binary form, and

FM 94 BUFR - Binary Universal Form for the Representation of meteorological data.

GRIB is designed for highly efficient encoding of regular distributed data. BUFR has a more universal nature and would need a more complex data description for those data which could also be represented in GRIB.

Both codes are designed to represent any meteorological data in a logic, efficient way. Binary representation, independent of any particular computer is employed. Each GRIB or BUFR message is a string of binary numbers, independent of any computer physical record, block, or file structure and includes both data and data description.

Everything seems to be simple and far away from any problem if GRIB and BUFR would be used worldwide. However, the use of GRIB and BUFR will also cause some problems which have to be discussed.

3. EXCHANGE OF METEOROLOGICAL DATA

Meteorological data may be exchanged through a system of distributed data bases or through the Global Telecommunication System (GTS).

A system of distributed data bases may be used to exchange data in the format of meteorological bulletins containing either binary or character data. Data may e. g. be stored and retrieved through a list of abbreviated headings. Recompilation of bulletins would require a list of contents of bulletins as well. Distributed data bases will be part of the GTS in the future but they will not replace the present form of the GTS within a few years.

Within the GTS data are transmitted as meteorological messages consisting of starting line, abbreviated heading, bulletin text and end-of-message signals. If the text part in a meteorological message contains only readable data both the starting line and the end-of-message signals are unique in the meteorological message. However, bit strings like starting line or end-of-message signals may appear anywhere in the text if a meteorological message would contain binary data. At present only the full data communication protocol as specified in CCITT Recommendation X.25 (including packet level) allows transmission of binary data.

The meteorological telecommunication networks are far away from the point that all of them are upgraded to handle X.25. Therefore, we have to choose between the possibilities of disconnecting all centres not connected by X.25 lines with the MTN from getting any binary data, of converting binary data to character codes at any Regional Telecommunication Hub (RTH) being connected by X.25 lines and being supposed to transmit data through non-X.25 lines or of transmitting data both in binary and character code on the MTN.

The WMO Commission for Basic Systems at its ninth session agreed that

> National Meteorological Centres should be encouraged to implement manual decoding and plotting of selected GRID products at the earliest opportunity to gain experience in advance of the implementation of computerized procedures.

Therefore, we have to discuss only the second and the third possibility. Both possibilities will reduce the speed of transmission. Anyway, the time for conversion of binary data to character codes at a RTH has to be compared with the additional time for the transmission of the data in character code.

4. LAYOUT OF DATA BASE SYSTEMS FOR BINARY INFORMATION

If meteorological bulletins are to be stored in a data base, there is no principal difference between storing binary or character data. Both types may be stored in the same way using e. g. a data descriptor like the abbreviated heading to perform quick queries. Therefore, there is no really need to go into the details of such a data base at this point.

However, if there is need for a data base assuring all kinds of quick access to specific data in the data base like

retrieval of a small number of single station reports,

retrieval of all reports from stations with the same WMO block number,

retrieval of all reports from stations located in a predefined geographical area,

we will need a different data base design for BUFR reports. On the other hand the text of meteorological messages containing data in GRIB code may be stored without any change and retrieved efficiently by a data base.

BUFR code permits a nearly unlimited variety of different data formats. If BUFR code would be stored as received, each retrieval of data from the data base would require a rather complex evaluation of the format of each BUFR report. Therefore, either

fixed predefined and international agreed subsets of BUFR would have to be used or

all incoming BUFR data will have to be converted to a limited number of locally defined BUFR formats.

Both possibilities would allow predefinition of the complete description of data formats in the data base. Storage and retrieval and the expansion of the BUFR formats could be managed in a very effective way. The exclusive use of predefined and internationally agreed data formats would restrict the universality of BUFR. However, e. g. for satellite data being prepared by a very limited number of centres the use of internationally agreed data formats could be suitable.

Anyway, the use of locally defined BUFR formats for the storage of data in a data base is more general.

Data flow control and error checking information could be easily included.

Only information which is really needed has to be stored in the data base.

BUFR messages from the GTS containing several reports from various observations could easily be split into single reports before the data are stored in the data base.

In-house made data bases are very much more efficient (factor 5 to 30 for both time and space) than commercially obtained data base management systems as was shown by benchmark tests at some centres

in the United States and in Europe. Therefore, the first session of the CBS Working Group on Data Management Sub-group on Data Representation which was held in Geneva in May 1989 stated that

> while a commercial DBMS could provide increased sharing of data resources, improved data security and integrity, increased portability of applications which access the DB, its consumption of hardware resources is still much too high. Therefore, in-home-development of data storage and accessing routines is still necessary and justifield.

The reports data base of the German Weather Service (DWD) will be used as an example for the following. This reports data base is prepared for the quasi-real-time storage of all types of incoming reports, both national, regional and global in BUFR-like formats. Incoming data (at present there are only character coded data) are converted into predefined BUFR formats by a separate decoding program. Incoming messages containing data from more than one geographical location in one single report (e. g. reports in FM 88 - SATOB) are splitted into separate BUFR observations. Our reports data base has no immediate telecommunication functions. Therefore, telecommunication formats will be prepared by a separate program, if necessary.

Data of different kind are stored in separate data bases. Each data base is divided in time boxes and the whole geographical area of each data base is divided in sub-areas. The general layout and contents of each data base are described by an architecture description file containing general information like data type, and extension of the data base in space and time. For each time box there is one administration data file containing a list of contents and address pointers to the data and one meteorological data file corresponding to each other. This structure allows quick access to data from a fixed time and a predefined area as well as access to any single report in the data base. If an administration data file would have been destroyed for any reason the corresponding meteorological data file of the data base could simply be used as input data for the recreation of the data base because input and internal data files have the same format.

5. THE FORMAT OF DATA IN THE DATA BASE OF THE DWD

In the data base of the DWD field data are stored as GRIB messages and reports are stored in a BUFR format. GRIB messages are stored as described in the WMO Manual on Codes, however, BUFR is used in an abbreviated form.

In order to simplify the retrieval, each observation is stored separately in the data base. In principle, we would need the overhead of one complete BUFR report for each single observation. Therefore, we are reducing the overhead in the data base by substituting BUFR Section 0 to 3 by a 16 octets information header containing all information which is not included in the data section of our BUFR formats but is needed to prepare a full BUFR (as described in the Manual on Codes). In addition the information header contains an identification number (different e. g. for manual and automatic observations, land and sea stations although the data format might be the same) and the pointer to the data base file, both being for local use only. Each data format in the data base is defined by one single table reference number in the local BUFR table D.

The information in the header combined with information in the data part of our data base formats would be sufficient to prepare a complete BUFR report.

The data part does not include the length of the data section. It starts with octet No. 5 of BUFR Section 4. The description of the data base BUFR formats is kept in a so called BUFR format table including some additional information facilitating the expansion of information from the internal data base format. The BUFR format table contains a data width in characters which can be used if BUFR data are to be converted into a character format. Such a character format could be used to create an output format in tabular form or the input for a character oriented data base. The BUFR format table contains also the starting positions of each data field and so called mnemonic abbreviations for each data field, the use of which will be described later. The starting position of a data field is in bits from the beginning of the data part.

The data part of each BUFR report starts with a report header. The format of the header is unique through all observational data formats, therefore, facilitating data base storage and retrieval. The report header contains decoding data, geographical location, date and time of validity, and identifier name or number of the station, platform or satellite. The decoding data were found very useful for both management (e. g. for the definition of cut-off times) and programmers of the decoding program (e. g. for tracing errors in reports and/or program).

The information of BUFR format tables is used both to prepare BUFR formats by the decoding program and to expand values from BUFR formats to integer and real numbers or to character strings by a service program. The same service program is used by the data base to extract the information for the administration data file from incoming data sets.

6. RETRIEVAL OF DATA

Data may be retrieved from the data base by data type, time of observation and geographical location (e. g. area on the globe or station identifiers like WMO block and station number, ICAO location identifier, SHIP call sign etc.). Retrieved information are written to a file which might be used by a post-processing routine to extract single values from the data. This post-processing routine is using the mnemonic abbreviations as defined in the BUFR format table to provide specific values from a report both in integer and real format or as a character string.

7. EXPANSION OF VALUES FROM BINARY FORMATS

GRIB code may contain either grid point data or spectral coefficients. Grid point data may simply be expanded to a two dimensional data array which could be used as input to the final output program. Spectral coefficients could be expanded either to grid point values as before or to an array containing the spectral coefficients itself depending on the need of the final output program.

The predefined data formats used by the data base of the DWD are described in the so called BUFR format table. The user of the expansion program may request expansion of specific values by a list containing the mnemonic abbreviations for each element. The starting position of specific values in the BUFR format table then allows direct access to each requested value. In addition simple conversion of data like e. g. from Kelvin to degrees Celsius could be provided. The mnemonic abbreviations are, if possible, derived from the symbolic letters used in the WMO manual on Codes. They are prepared for direct use as names of variables whithin a FORTRAN program and they are directly connected with reference

numbers in BUFR tables B, C and D. If a specific reference number appears more than once in a given BUFR format the mnemonic abbreviation is made unique by appending an additional character.

The expanded values could then easily be used as input for all kinds of presentation programs like plotting charts or preparing data tables.

8. FINAL REMARKS

Binary data should in most cases be pre-processed before they are ready for storage in a data base. With preference the data base should be an in-house made one which may be optimized to solve the specific problems at any centre. Data should be retrieved from the Data base as whole reports and a post-processing routine should be used to extract single values from the data sets.