Deutscher Wetterdienst
Wetter und Klima aus einer Hand

IN23D-0901 - Addressing the ‘R’ of the FAIR Data Principles with BUFR

Markus Heene, Deutscher Wetterdienst (DWD)

Meteorological Codes

The WMO (World Meteorological Organization) has more than 190 members. These members exchange their observations and measurements around-the-clock. The Manual on Codes contains the international codes for these meteorological data. The Traditional Alphanumeric Codes (like SYNOP FM 12) are no longer maintained and the migration to BUFR is ongoing. BUFR is a table driven code form which means you can easily extend it without changing the software. BUFR is maintained by an WMO Expert team.

Message Decoding

DWD collects in its role as GISC all global available observations and measurements intended for global exchange and additional data from other networks. These data are encoded in different formats. In a first step all these different formats are processed by our decoding system. The software decodes the different formats and than maps the entities onto internal harmonized BUFR templates. The resulting BUFR is used for storage and further processing. Figure 1 shows this process for synoptic messages. You will likely find similar processes in all centers which run a global model.

Complexity Reduction

The previous step reduces the complexity of the succeeding steps. Each of the following processes needs only to decode the relevant harmonized BUFR template(s) and not the full variety of all incoming formats.

In addition it reduces the number of decoding software we need to provide and maintain (in particular for historical formats).

Furthermore the likelihood of misinterpretation is significant reduced due to the precise description of the elements in the BUFR tables.

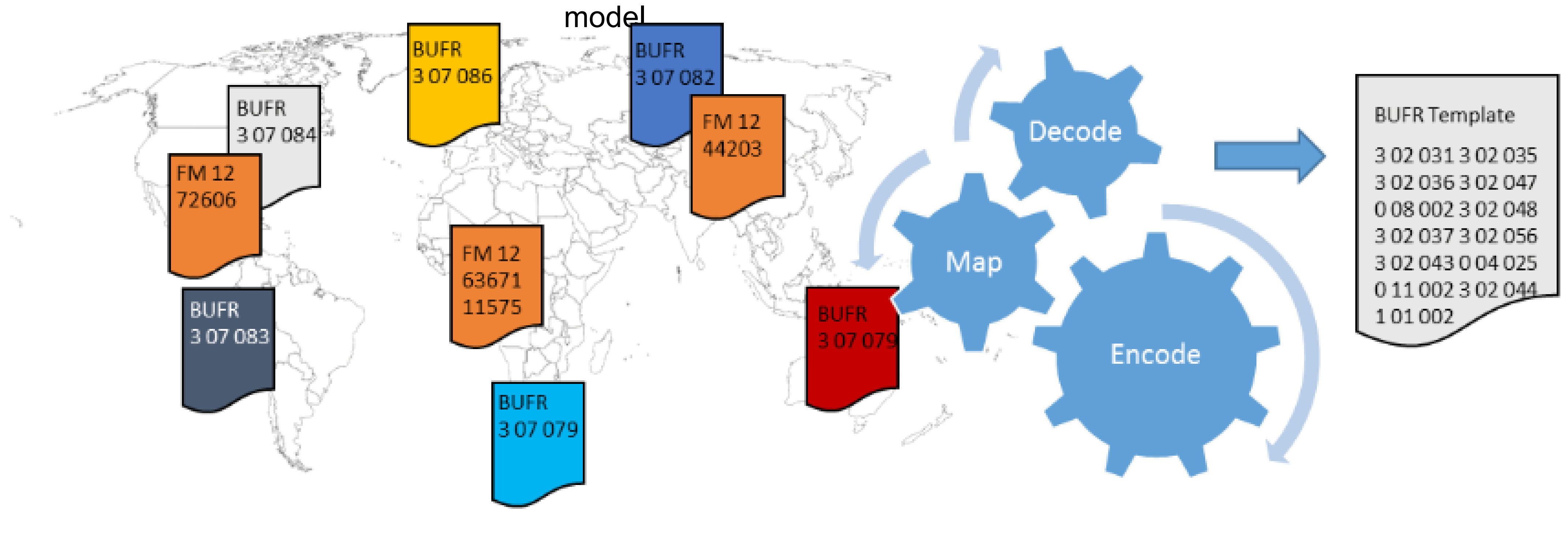


Figure 1: Schematic overview of the collection and processing of synoptic observations

BUFR as an archive format

DWD maintains an „eternal“ memory of all observations and measurements. All these data are stored in a relational database. An ETL process decodes the BUFR and loads the values into the appropriated tables. While the latter database is mainly used by climatologist the Data Management System is mainly used by the numerical weather prediction (NWP). Figure 2 shows the schematic overview of the data storage process.

In our case we use BUFR as an uniform transport and storage format.

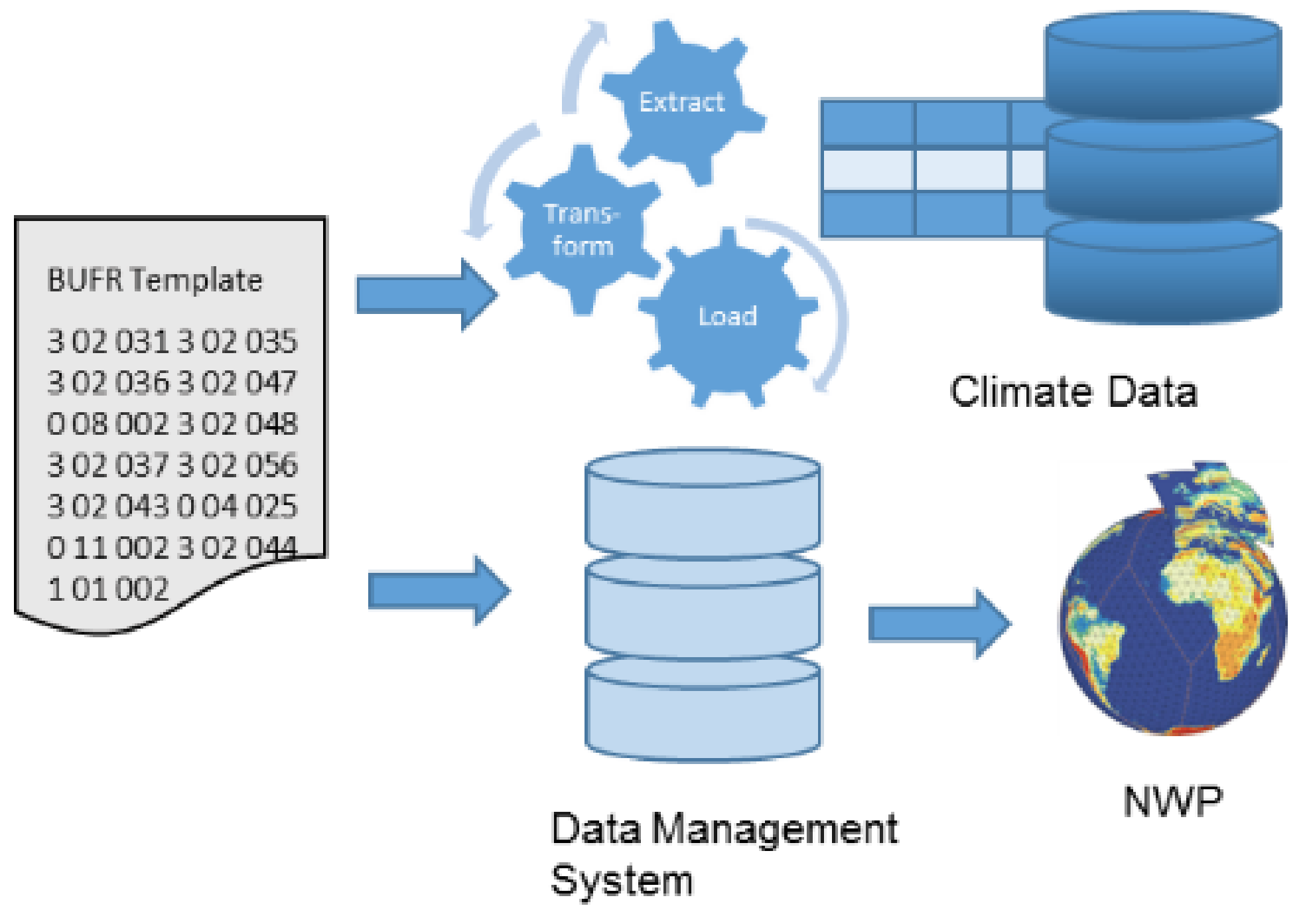


Figure 2: Schematic overview of the data storage process

BUFR Grammar - Introduction

BUFR consists of sections (see figure 3).

Section 3 “Data description section” is the blue print for Section 4 “Data section”. Section 3 consists of descriptors. A descriptor contains 3 parts: F (2 bits), X (6 bits) and Y (8 bits).

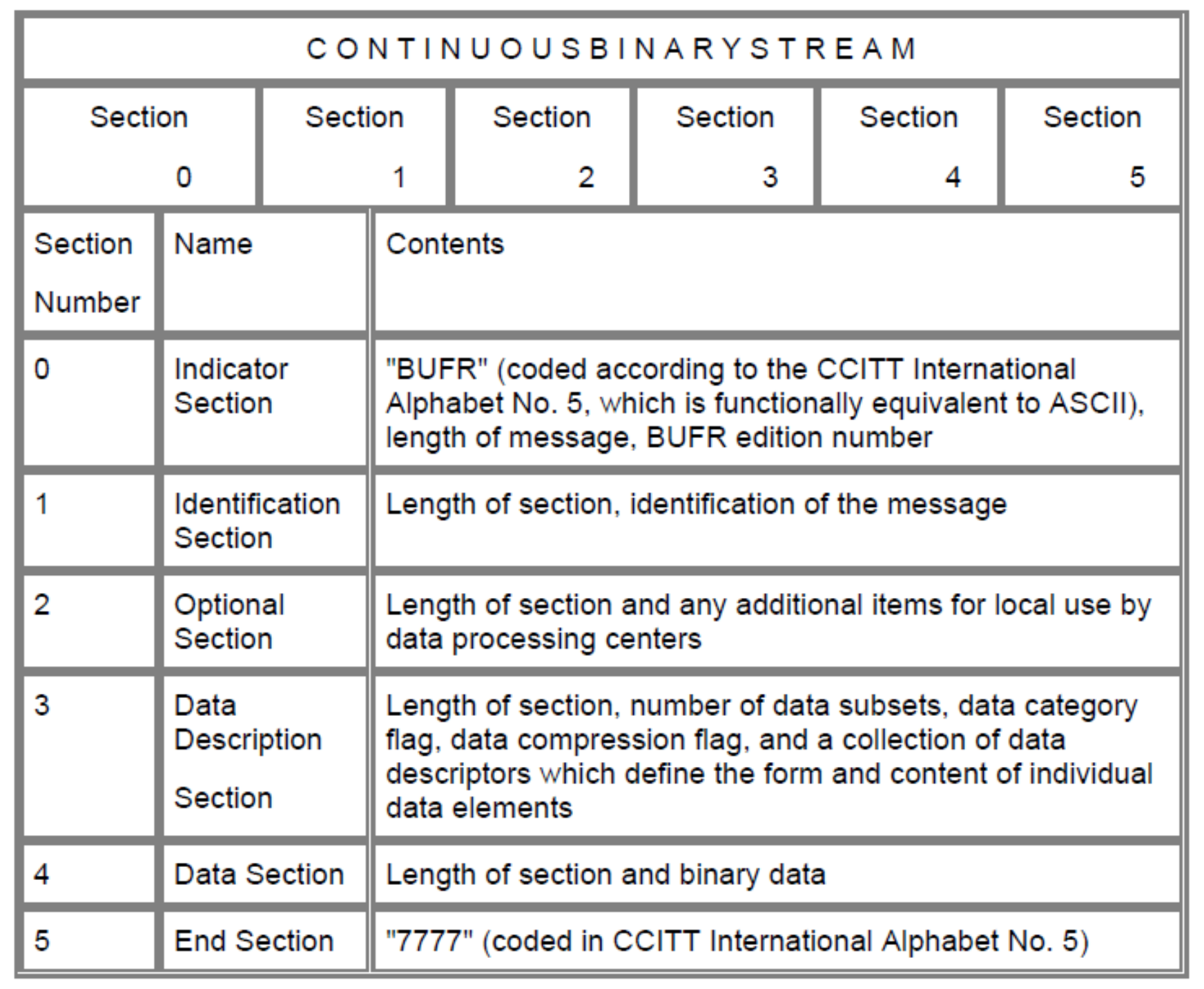
F = 0 is the "element descriptor" which defines a single data item (temperature, pressure, humidity, ...)

F = 1 is the "replication descriptor". In this case X indicates the number of descriptors to be repeated, and Y the total number of replications. For Y = 0 we call the descriptor "delayed replication descriptor" and the number of replications is encoded in the data section with the so called "delayed descriptor replication factor" e.g. 0 31 000, 0 31 001, ...

F = 2 is the "operator descriptor".

F = 3 is the "sequence descriptor". A sequence descriptor defines a list of element descriptors, replication descriptors, operator descriptors and/or sequence descriptors.

For the construction of Section 3 with the 4 types of descriptors (element, replication, operator and sequence descriptor) described above certain rules apply. The rules are described in detail in the Manual on Codes starting at regulations 94.1 and following.



CONTINUOUSBINARYSTREAM					
Section 0	Section 1	Section 2	Section 3	Section 4	Section 5

Section Number	Name	Contents
0	Indicator Section	"BUFR" (coded according to the CCITT International Alphabet No. 5, which is functionally equivalent to ASCII), length of message, BUFR edition number
1	Identification Section	Length of section, identification of the message
2	Optional Section	Length of section and any additional items for local use by data processing centers
3	Data Description Section	Length of section, number of data subsets, data category flag, data compression flag, and a collection of data descriptors which define the form and content of individual data elements
4	Data Section	Length of section and binary data
5	End Section	"7777" (coded in CCITT International Alphabet No. 5)

Figure 3: BUFR sections

BUFR Grammar – Implementation

Looking from a different perspective the rules as a whole can be interpreted as a language. This language is a set of valid sentences, and sentences are composed of phrases and clauses. Hereby a sentence structure follows a grammar. In our case the descriptors represent the phrases and clauses. The following sequence of descriptors is a valid sentence: 3 10 014 2 22 000 2 36 000 1 01 103 0 31 031 0 01 031. The rules for applying and combining the descriptors are our grammar. Additional rules for the replication descriptors are implemented as listener pattern.

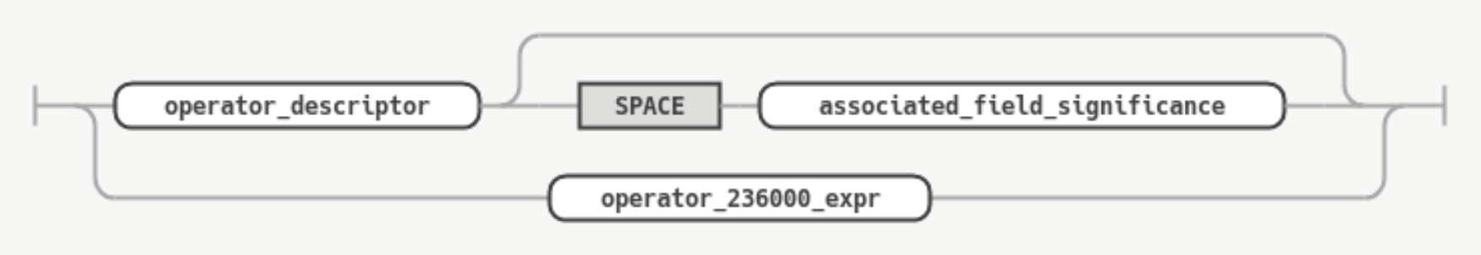
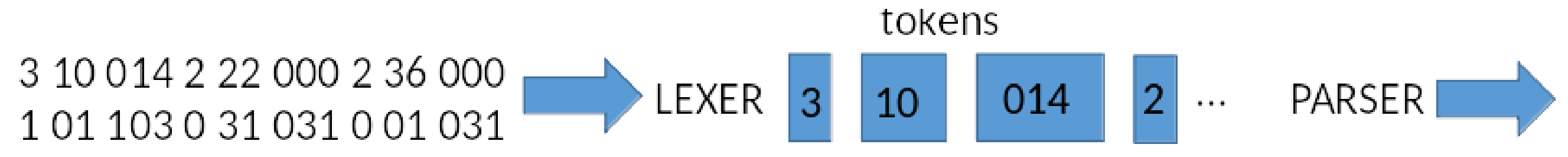


Figure 4: Railroad diagram for „operator descriptor expr“

Figure 4 shows the railroad diagram for the parser rule “operator descriptor expr”.

Figure 5 shows the data flow on a concrete example. The lexer analyses the input template and breaks it down into tokens. In a second step the parser recognizes the sentence structure based on our grammar. The result is a parse tree.



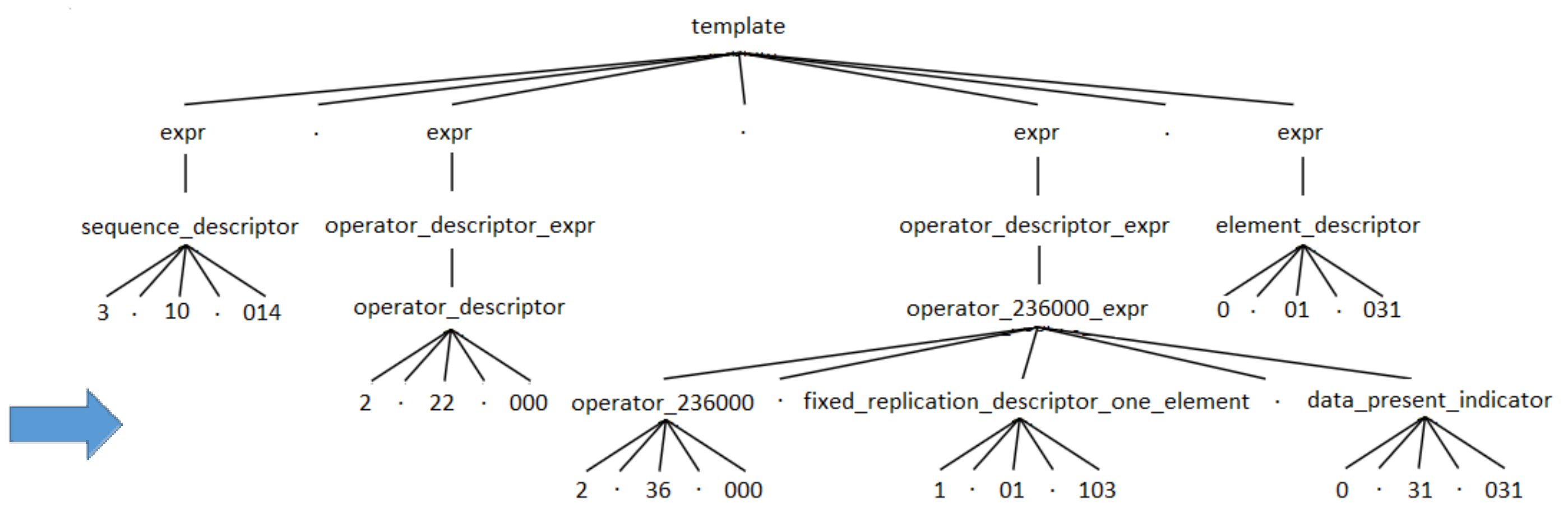


Figure 5: Basic data flow of a language recognizer

A basic version of a BUFR grammar is available on github[1]. The grammar is formulated in ANTLR[2]. With the help of the grammar a BUFR template developer can now check her/his template for syntax errors. The BUFR specification is maintained by WMO as a human readable document while a machine readable grammar does not exists. Furthermore no reference implementation of the BUFR specification exists. Therefore a validation of a BUFR decoder/encoder or an BUFR against the specification isn’t easy. This step could be simplified with a grammar.

Outlook

It is planned to present to IPET-CM the grammar next year and to initiate a discussion about the usage of a grammar. Furthermore the grammar is intended as a contribution to the upcoming development of BUFR 5.

Resources:
[1] https://github.com/mheene/bufr-grammar
[2] https://www.antlr.org

Abteilung TI 1 Systeme und Betrieb, Referat TI 12 Datenmanagement
Contact: Markus Heene, markus.heene@dwd.de