CONVERSION MODEL FOR OPEN METEOROLOGICAL DATA

Tran Thị Tien¹, Trieu Thanh Ngoan², Ngo Minh Phuoc¹

¹Information Communication Technology Center, Ministry of Science and Technology
²College of Information and Communication Technology, Can Tho University

ttien@most.gov.vn, ttngoan@cit.ctu.edu.vn, phuocnm@most.gov.vn

ABSTRACT: Data is considered as an asset of the global economy. Along with the term "open", Open data becomes the new trend of agencies, organizations, and governments. Providing Open Data helps develop businesses and startups since the experience of countries around the world has a series of startups born from Open Data warehouses. Currently, Vietnam is in the early stages of building an open database, which includes meteorological data. This data is important since it helps to process and determine future climate expectations. BUFR/SYNOP are the common formats that are used by almost all meteorological services in the world. Those types of formats cannot be used for open purposes since they are not ensuring the accessibility feature of open data. In this paper we propose a new data model for meteorological data to convert weather data into machine-readable formats. The experiments will be done in converted data for weather forecast and explanations.

Keywords: Open Data, Weather Forecast, BUFR/Express, Environment Simulations.

I. INTRODUCTION

The term Open data [1][2] has been used for a long time and is becoming popular in many fields such as science, technology, law, and medicine. It represents the trend of publicizing, sharing information and data so that people can access, reuse, and create value-added applications and services from this raw data. Recently, many governments around the world are interested in Open Data. It becomes a trend of developed countries and becomes an indicator of the level of e-Government development of the United Nations¹ and many international organizations. Some governments have completed the construction of digital data centers and started publishing data. Vietnam is in the early stages of digitizing and opening information of the government [3]. It is a long-term program, requiring concentrated investment and reasonable solutions suitable to the specific conditions of the country. However, Vietnam is also enjoying certain advantages when people are gradually becoming aware of using technology in general and using the Internet in particular to update and capture information.

Weather data is particularly important, as everyone from individuals to huge companies can benefit from it. Applications that make use of open weather data can therefore potentially have a huge impact. The weather data collected every day can have a range of benefits if it is made open and interpreted in the right way. The crucial measurements taken every day can be used in different ways by people from all walks of life if the right tools are provided to them. In this way, simple information can create new benefits for everyone. Currently, most meteorological observational data are already being produced and transmitted in BUFR [5], [7], [8], and according to WMO’s migration plan (World Meteorological Organization ²), in the near future all meteorological observational data should be coded in BUFR. BUFR has many good characteristics that make it a meteorological code suitable for the 21st century. However, this data format cannot use for the open purposes since it is not easy use by everyone.

Although data published in any format associated with an Open License may be considered Open Data, the type of data format used can have important implications for usability of the data. In this paper, we propose a new data model for converting meteorological data from BUFR format to other common formats, such as CSV, JSON, and XML. The experiments were conducted on weather raw data from several weather stations in the world. We use the converted data with CSV format for other purposes, weather forecast and explanations.

The structure of this paper is organized as follows. Section 2 presents the background about open data. Section 3 provides a data model for meteorological data with BUFR/Express. Section 4 will be an experiment where we use our data model for reading and processing weather data. The conclusion of the paper is presented in section 5.

II. OPEN DATA

Open data is data that can be freely used, reused and redistributed by anyone. An important feature of Open Data is its accessibility and readability to use for machine-readable requests. The main features of Open Data are:

- Availability and accessibility: Data must be available and can be downloaded over the Internet. Data should also be readily available in a convenient and standard format.
- Reuse and Redistribution: Such data must be provided under terms that permit reuse and redistribution.
- Globality: There is no distinction between groups of people in using data. For example, there are no "non-commercial" restrictions that obvstruct ‘commercial’ use or restrictions on use for certain purposes.

¹ https://www.un.org/en/
² https://www.wmo.int
In recent years, the concept of open data, especially open government data, has become a central role in the way information is collected in urban centers. When government data are made accessible and re-usable, they enable individuals, organizations and even governments themselves to innovate and collaborate in new ways.

The benefits of open government data:

- **Transparency**: Open Data makes it easier to monitor government activities. Thus it can support public oversight of governments and help reduce corruption by enabling transparency.
- **Public Service Improvement**: Citizens can use Open Data to contribute to public planning, or provide feedback to government ministries on service quality.
- **Innovation and economic value**: Businesses and companies are using Open Data to better understand potential markets. By making these data publicly available, it provides opportunities for individuals and businesses to utilize the data to create value-added products and services for commercial gain.

In contrast to traditional closed innovation approaches where the innovation lies within the boundary of an organization, open innovation accepts an open boundary between an organization and its surrounding environment. This allows organizations to make use of external knowledge in developing services. The datasets available in machine-readable format to facilitate application developers to tap on these data to develop innovative e-services that can be made available over the web or as mobile applications.

### III. METEOROLOGICAL DATA MODELING

In the last few years of the twenty century, weather radar became a highly important tool for meteorology, especially with regards to short term forecasting. Weather radars were quite expensive tools thus people try to obtain good coverage of the area of interest at minimum costs, which means having as few radars as possible. It raised the problems of transmission for sharing radar data. The BUFR was the result of expert meetings and periods of experimental usage by several meteorological data processing centers. Lately, it is used as a standard for radar data representation of sharing meteorological data.

#### A. Weather data

Weather data include any facts or numbers about the state of the atmosphere, including temperature, humidity, precipitation, wind speed, and pressure. Weather data can be represented as matrix $A_{m,n}$, in which $m$ is the number of observations recorded and $n$ is the number of features (temperature, humidity, etc.).

$$A(m, n) = \begin{pmatrix}
a_{1,1} & \cdots & a_{1,n} \\
\vdots & \ddots & \vdots \\
a_{m,1} & \cdots & a_{m,n}
\end{pmatrix}$$

Each observation is the information recorded about the state of the weather and the place that records the information. In other words, each line of a weather data file after the header row represents one observation of the weather by one weather station at a specific time. Each column represents a feature of the atmosphere.

Nowadays, people have some amazing ways to collect these kinds of data from all sorts of places: the ground, the air, and even from the space. Meteorologists use all kinds of high-tech equipment (such as weather station instruments, radar systems, and satellite images) to measure the weather. Weather station [6] has many kinds of instruments for measuring different characteristics of the weather on the ground. For example, a thermometer is used to measure the temperature by allowing mercury inside the thermometer to expand as it gets hotter and contract as it gets cooler. A scale on the outside of the thermometer will match up the change of the air temperature. Barometer is another example used to measure air pressure. This is also an important measurement for weather forecasts. A change in air pressure indicates a change in weather is coming. Normally, if there is a rise in air pressure, a clear sky can be expected and if there is a fall in air pressure, it is likely to have storm clouds.

Radar stands for Radio Detection and Ranging. It is another tool for meteorologists gathering weather data. In recent years weather radar has become a highly important tool for meteorology, especially with regards to short term forecasting. Weather radars are quite expensive tools thus people try to obtain a good coverage of the area of interest at minimum costs, which means to have as few radars as possible. The satellite has become an important source of information on weather data since it was first introduced. The satellite can record long-term problems such as polar ice melting.

Weather data can be retrieved from multiple sources thus it brings the diversity and complexity for data integration, which is combining information from various sources. The combination of all sources with geo-location information can fill geospatial gaps in weather data for simulations and predictions. The WMO has provided several kinds of code forms for ease of weather data transmission thus there are systems and networks of regional and global weather data exchange. BUFR is a commonly used standard for sharing meteorological data.
B. BUFR standard

BUFR (Binary Universal Form for the Representation of meteorological data) is the WMO code form that is designed to represent a continuous binary stream of any meteorological data. The core concept of BUFR standard is self-descriptive nature, which helps this standard in accommodating changes. It only needs to have additional data description tables when there is new observation data. The data description is a major part of the BUFR standard documentation.

A BUFR message contains any kind of observational data and a complete data description of the data. The term “message” means that the BUFR standard is used as a data transmission format. A BUFR message is a continuous binary stream comprising of six sections (Example as in figure 1). Each section is made up of a series of octets, coined to qualify one byte as an 8-bit sequence. In theory, there is no upper limit to the size of a BUFR message. However, by convention, BUFR messages are restricted to around 15,000 octets.

The sections of a BUFR message is organized as follows:

- Section 0 is called indicator section containing 8 octets, with the string "BUFR" indicating the start of a message. The BUFR edition number is also provided in section 0.
- Section 1 is identification section containing some information about the message such as the master tables version used to decode the message.
- Section 2 is an optional section for any additional uses.
- Section 3 is the data description section, which is a list of descriptors. It includes a list of data elements that are contained in each data subset.
- Section 4 is the actual data specified by section 3 with 2 formats: compressed or uncompressed.
- Section 5 is the character string "7777" indicating the end of the message.

The BUFR is a strong and complex binary format with self-descriptive nature. Thus there is a need of a model for conversion a BUFR message into other formats.

C. Express Modeling Language

Express is a standard data modeling language, which is formalized in the ISO standard 10303-11 [10] (within the STEP - Standard for The Exchange of Product model data). It is an object-flavored lexical language that is firstly designed to represent the models of industrial products. The data modeling language helps define data objects and the relationships between the objects and enabling the exchange of data between the objects. A data model can be defined in two ways, textual or graphical. In a textual form, a SCHEMA is clarified in which various data types and the structural constraints and algorithmic rules can be defined. The Express- G is the graphical representation for all details formulated in the textual form. The advantage of using Express-G is that the information can be presented more understandably.

Express provides a series of data types for building blocks in a schema. The most important data type in Express is the entity data type. Entity attributes can relate an entity with other entities. In brief, Express can be used to model data and data relationships with a general inheritance mechanism and can be used as a procedural programming language to specify constraints on data instances.
**D. BUFR/Express Meteorological Data Modeling**

A data modeling process is necessary to define and analyze the BUFR data. It not only defines the data elements, but it also defines the structures and relationships between the elements. This study uses Express as a test case for meteorological data modeling. Express modeling language will help to analyze the meteorological data in BUFR messages and provide an accurate BUFR parser. A data model is defined for BUFR messages using Express modeling language. As an example, the descriptor entity in our model is represented as follows:

```
ENTITY descriptor;
  fxy : STRING;
  f,x,y : INTEGER;
  element_name : STRING;
  note : STRING;
  unit : STRING;
  scale : STRING;
  reference_number : STRING;
  bitwidth : INTEGER;
END ENTITY;
```

The textual representation of this schema is important as an input for SDAI (Standard data access interface). SDAI is an abstract specification on how to deal with Express schema and can be mapped to various programming languages. This paper uses JSDAI [11], which is a Java application-programming interface for reading, compiling, and writing object-oriented data defined by an Express model. The java code with classes of objects is auto-generated with JSDAI API. It is further added some functions for reading raw binary data and export the result into commonly used and machine-readable formats – CSV, JSON, and XML.

Section 3 of BUFR contains all descriptors necessary for defining and representing data in a message. This is the important information to convert a BUFR message into other formats. This data description is all contained in tables [4] which are the major part of the BUFR documentation. We build up a tree structure for each BUFR message descriptors (an example as in figure 2).

![Figure 2. An example of a BUFR message descriptors in tree structure](image)

We use depth-first search [18] for traversing tree structures. This algorithm starts at the root node and explores as far as possible along each branch before backtracking. The last task is to match the descriptors with the bit stream, and extract the values out of the stream.

**IV. EXPERIMENTS**

Weather forecasting is the task to predict the state of the atmosphere in a given location. In the past, the weather forecast has been done through physical models of the atmosphere as a fluid. It becomes the problem of solving sophisticated equations of fluid dynamics. In recent years, machine learning algorithms have been used to speed up weather data modeling, a computationally intensive task. Machine learning algorithms learn from data and produce relevant predictions. In addition to prediction, there is a need of providing knowledge about domain relationships inside the data. Interpretable machine learning is the use of machine learning models for the extraction of knowledge in the data. This paper will used the converted data in CSV format for the problem of weather forecast and explanations. An illustration is shown on characteristic variables of meteorological data.

**A. Data Description**

The experiments use data collected from Meteomanz.com [12], an FTP server of the National Oceanic and Atmospheric Administration (NOAA). This database consists of 13,000 weather stations all over the world specifying by geographical locations. It is noted that the meteorological stations provided are all registered with WMO and each will receive a 5-digit WMO index for identification. The weather data in this database is from the year 2000 until the present.

The data used are hourly collected in a year (from June 2019 to May 2020) in three different characteristic regions. The first dataset is retrieved from Rachgia Vietnam (station number 48907), which belongs to a tropical region. The second dataset is from Brest France (station number 07110), which is inside the temperate zone. The last one is data from Kemi Finland (station number 02864), which has a continental climate with freezing winters and mild
summers. Each record is a BUFR message with a list of descriptors and the data section corresponding with the descriptors.

B. Data conversion

We convert the BUFR messages into CSV files corresponding to the datasets. The CSV data are as follows:

- Rachgia, Vietnam: 1.271 records of 121 columns. There is a lack of data recorded in this city, especially with no data for the whole month in October 2019.
- Brest, France: 7.813 records of 191 columns. The weather data in this city is more complex that has different data descriptors from one day to another day.
- Kemi, Finland: 8530 records of 109 columns. Data recorded in this city is simple with fewer descriptors.

Data cleaning is a critically important step for any machine learning algorithm. In the dataset, there are columns without data (indicate as missing value). Columns that contain a single value or columns with “missing” indication are referred to as zero-variance predictors [13]. It means that measuring the variance of these predictors will return a zero. These columns do not contain any information for modeling. It is simply to remove the zero-variance predictors from the dataset.

C. Weather forecast and explanations

In recent years, machine learning algorithms have been widely used for intelligent weather prediction since it is not necessary to have a complete understanding of complex processes that govern the atmosphere. In this study, we use random forest approach [9], a combination of a large number of decision trees, for building prediction models. A randomForest [15] is a R package that provides methods to create models. The most common outcome of the decision trees is used as the final output. Take an example, we build up the model and predict the Horizontal Visibility knowing the other features such as temperature, humidity, and pressure. The prediction for a record in the test data is 10.5132 while the real value is 10. There are available problems about the interpretability of machine learning models for weather prediction.

IML [14] is a R package that provides methods making machine learning models interpretable. Most methods in the package have been implemented in other packages, such as Feature importance [16], Interaction effects [17], and Shapley value [21]. However, this IML package puts all methods in one place thus it is more convenient with the same syntax and consistent functionality.

1. Global interpretation

Measuring how the importance of each feature [16] for the predictions will help us to understand how the models make predictions based on the features and their influence on the underlying model structure. As an example, we measure the importance of features for predicting Horizontal visibility (figure 3). In the three places, Relative humidity is an important feature affecting the predictions. The Cloud amount and Height of the cloud are also important features for estimating the Horizontal visibility.

2. Local interpretation

Shapley value [19][20] is used to explain a specific prediction, which is the contribution of each feature to the difference between actual prediction and average prediction. In figure 4, the actual prediction value is 40.47, which is 3.18 below the average prediction of 43.65. The air temperature has the most positive contribution and the cloud
amount has the most negative contribution. The sum of Shapley values yields the difference between actual and average prediction.

Figure 4. Shapley values of features on predicting Horizontal Visibility, Rachgia, Vietnam

Figure 5 is the scatter plot for feature importance scores with Shapley values of 100 random predictions on Horizontal Visibility in Kemi, Finland. In the whole feature space, it is clear that the Relative Humidity is a confusing feature since its values have the most positive and negative contributions to the prediction outcomes. Figure 6 is the scatter plot for the Shapley values of 100 random predictions on Horizontal visibility in Brest, France.

Figure 5. Shapley values of 100 random predictions, Kemi, Finland

Figure 6. Shapley values of 100 random predictions, Brest, France
V. CONCLUSION

Many meteorological stations are provided for collecting and sharing weather data. It provides a good basis for weather simulations and disaster forecasting based on multiple data sources. This study provides a data model for weather data to convert from BUFR to other common formats for the open purposes. The common formats have a lot of implications for the user and one standard cannot fulfill the full list of needs that arise in the countless circumstances in which open data is used. The experiments have been conducted with our proposed model. The results of this study may initially support the creation of an open weather database in Vietnam. This later leading to the improvement of meteorological data understanding and deciding weather feature weights for the more accurate predictions. The ultimate goal of all researches is to find which counteractions need to be done for the best effects.

REFERENCES

MÔ HÌNH ĐỞI LIỆU CHO ĐỮ LIỆU KHÍ TƯỢNG MỞ
Trần Thị Tiến, Triệu Thanh Ngoan, Ngô Minh Phước