For this assignment, we will be using the METAR weather report dataset provided by Aviation organization. METAR is a format for reporting [weather](https://en.wikipedia.org/wiki/Weather) information. A METAR weather report is predominantly used by [pilots](https://en.wikipedia.org/wiki/Aviator) in fulfillment of a part of a pre-flight weather briefing, and by [meteorologists](https://en.wikipedia.org/wiki/Meteorology), who use aggregated METAR information to assist in [weather forecasting](https://en.wikipedia.org/wiki/Weather_forecasting).

Raw METAR is the most common format in the world for the transmission of observational weather data. It is highly standardized through the [International Civil Aviation Organization](https://en.wikipedia.org/wiki/International_Civil_Aviation_Organization)(ICAO), which allows it to be understood throughout most of the world. We have got the raw data from Metar website tgftp.nws.noaa.gov. For example, the following URL will download the most current weather observations for the Cadillac / Wexford County:

<http://tgftp.nws.noaa.gov/data/observations/metar/decoded/KCAD.TXT>

Although the data is updated on an hourly basis, we'll be working with only two datasets and convert them into XML representation to make it more readable. We will also create a DTD for the data.

The general structure of the DTD for Metar is below. Please refer the attached XML and raw data files to get more information.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parent** | **Element** | **Attribute** | **Description** | **Type** |
|  | <METARS> |  | This is the root element of the DTD. This can have Metar as sub elements. | Once and Only Once |
| <METARS> | <METAR> |  | This element is to provide the information about Metar observation | Zero or More Occurrences |
|  |  | ICAO-LOCATION | This four-letter code is called the "ICAO (International Civil Aviation Organization) location, | Required |
|  |  | UTC-DATE | The UTC date when the reading was taken | Required |
|  |  | UTC-TIME | The UTC time when the reading was taken | Required |
|  |  | CYCLE | Number of cycle the reading was taken | Required |
| <METAR> | <LOCATION> |  | Location details of the reading, it is divided in Political and geographical location information | Zero or One Occurrences |
| <LOCATION> | <POLITICAL> |  | This element provides the information about location from political perspective | Zero or One Occurrences |
| <POLITICAL> | <STATION> |  | Airport information if available. The element value is of PCDATA type. | Zero or One Occurrences |
| <POLITICAL> | <CITY> |  | City of airport. The element value is of PCDATA type. | Zero or One Occurrences |
| <POLITICAL> | <STATE> |  | State of airport. The element value is of PCDATA type. | Zero or One Occurrences |
| <POLITICAL> | <COUNTRY> |  | Country of airport. The element value is of PCDATA type. | Zero or One Occurrences |
| <LOCATION> | <GEOGRAPHICAL> |  | Geographical location of airport. The element value is of PCDATA type. | Zero or One Occurrences |
| <GEOGRAPHICAL> | <LONGITUDE> |  | Longitude of airport. The element value is of PCDATA type. | Zero or One Occurrences |
| <GEOGRAPHICAL> | <LATITUDE> |  | Latitude of airport. The element value is of PCDATA type. | Zero or One Occurrences |
| <GEOGRAPHICAL> | <ALTIMETER> |  | Height from sea level of airport. The element value is of PCDATA type. | Zero or One Occurrences |
| <METAR> | <LOCAL-DATE> |  | Local date at airport. The element value is of PCDATA type. | Zero or One Occurrences |
| <METAR> | <LOCAL-TIME> |  | Local time at airport. The element value is of PCDATA type. | Zero or One Occurrences |
|  |  | TIMEZONE | Time zone of the airport | Required |
| <METAR> | <WIND> |  | Wind speed. If wind is blowing the direction can also be provided using below attribute. The element value is of PCDATA type. | Zero or One Occurrences |
|  |  | DIRECTION | Direction of the wind. Its optional as sometimes there may be no wind. | Optional |
| <METAR> | <VISIBILITY> |  | How far the visibility is. The element value is of PCDATA type. | Zero or One Occurrences |
| <METAR> | <SKY-CONDITIONS> |  | Sky conditions it can be make as controlled. The values can be Cloudy, Overcast, clear etc. The element value is of PCDATA type. | One and Only one |
| <METAR> | <TEMPERATURE> |  | Temperature measurement given in Metar raw data in two units Fahrenheit and Celsius. We dropped the redundant information and will keep only one unit. The element value is of PCDATA type. | One and Only one |
|  |  | UNITS | Unit of temperature | Controlled (F|C), default F |
| <METAR> | <DEW-POINT> |  | Dew Point measurement given in Metar raw data in two units Fahrenheit and Celsius. We will keep the information in only one unit. The element value is of PCDATA type. | Zero or One Occurrences |
|  |  | UNITS | Unit of Dew Point | Controlled (F|C), default F |
| <METAR> | <RELATIVE-HUMIDITY> |  | Relative humidity in percentage. The element value is of PCDATA type. | Zero or One Occurrences |
| <METAR> | <ALTIMETER> |  | This is a barometer setting given in inches of mercury and hectopascals (hPa). We will keep information in only one units. The element value is of PCDATA type. | Zero or One Occurrences |
|  |  | UNITS | Units of Altimeter | Controlled (in.Hg|hPa) Default - in.Hg |
| <METAR> | <OB> |  | Observation. The element value is of PCDATA type. | Zero or One Occurrences |

* How did you decide to represent the data in the way that you did? Why did you choose the elements and attributes that you did?

As we know XML documents are composed of elements, so first we isolated the atomic elements from the raw data and then converted them to human readable format. We separated the presentation from the content and removed the redundant information. Also, added the structural elements in surrounding. For example – Added separate political and geographical locations and surrounded that by Location element, to make it more high level.

We also removed the redundant information as some of the information was provided in two units we only keep one.

We normalized the data and identified the unique keys in the form of attributes for the Metar observation, to identify every Metar reading as unique one. We used ICAO-Location, Time and Date and Cycle to identify the individual Metar observation.

We put the essential details/core content expressed in the XML in the form of elements. However, Attributes has been used for the meta data about the core information.

* What were the hardest decisions you had to make in this design process?

Below were the hardest decisions I had to make in the design process.

* When do I use elements and when do I use attributes for presenting bits of information?
* When do I require an order for elements, and when do I just allow arbitrary order?
* When do I use wrapper elements around sequences of similar elements?
* How does your DTD design support data independence?
* DTD design presented in this document physical data independence as using this our application programs, are able to refer to the data at the logical level, ignoring the details on how the data is physically stored and accessed by the system.
* Also, any changes to the DTD like changes to element structure, addition/removal of element or attribute from the DTD will not affect user views. They will see the data like before. This way we have logical data independence.
* How may your DTD design support the overarching goals of data curation (revisit objectives and activities of Week 1)?

We have designed a DTD which will support many of the goals of data curation.

* Collection – DTD will support the collection and acquisition of data.
* Organization - We have an appropriate DTD model with correct standards.
* Storage - The designed model supports reliable and effective storage.
* Preservation – The data is understandable without detailed knowledge.
* Discoverability - The designed model supports the ability to search for and locate relevant data.
* Access – Data retrieval is supported in the designed model.
* Workflow – The designed model supports the ability to systematize data workflows.
* Identification – The designed model supports the ability to identify, authenticate, and validate data.
* Provenance - The designed model supports identifying what inputs, processes, and calculations are responsible for data values.
* Modification - The designed model supports management of corrections and updates.
* Integration – The designed model supports integration of data from different sources.
* Reformatting – The designed model supports reformatting for use by different tools or to match new format standards.