EARTH SCIENCE MARKUP LANGUAGE:
A Solution to the Earth Science Data Format Heterogeneity Problem

Rahul Ramachandran[1] Helen Conover[1], Sara Graves[1] and Sundar Christopher[2]

Information Technology and Systems Center[1]
Department of Atmospheric Science[2]
University of Alabama in Huntsville
Huntsville, AL - 35899

1. INTRODUCTION

NASA's Earth Science Enterprise, formerly Mission to Planet Earth, was established to use the agency's advanced technology to understand and predict how the land, oceans and atmosphere interact as a system in influencing climate change. To accomplish this objective, NASA is flying a new series of satellites that take a variety of measurements important to Earth Science research. Terra, the first in this series was launched in December 1999, and began collecting science data in February 2000. Additional data will be taken from smaller, more narrowly focused satellite missions. Coupled with satellite data, NASA and other agencies continue to collect data from aircraft, ground-based and other space borne instrumentation to validate and document high quality Earth Science data. The unprecedented volume of data expected from these missions will be collected, processed, and distributed by NASA's Earth Observing System Data and Information System (EOSDIS). One of NASA's goals is to promote open access and use of this data by the general public, including the academic and industrial communities.

Computer models are also used to understand the physical processes and to predict and understand the science of global climate change. However, the Earth observation data sets used to initialize the models are heterogeneous in nature. Scientists are confounded by the various data types, formats and systems used for Earth Science data. For example regional atmospheric air pollution models have to deal with over 20 different data sets. These formats range from HDF to GRADS to GIS to MCIDAS to AIRS, each of which has a large manual to needed understand the structure and format of the data. It is almost impossible for individual scientists or even scientific groups to have expertise at their disposal to deal with each of these data types.

Even important new data is often not incorporated into models if it is in a format that is new to the scientists or their group. While considerable effort has gone into making Earth Science data usable and accessible to scientists, the myriad of formats, data types, navigational models etc. continue to stymie the use of more than a few data sets by individual scientists or groups. In this paper we discuss a new strategy for manipulating and analyzing Earth Science data sets using the Earth Science Markup Language (ESML) [1], [2]. A pilot project using ESML in a data analysis tool is currently underway, and will also be described.

2. ESML

ESML is an interchange technology developed by the Information Technology and Systems Center at the University of Alabama in Huntsville as a solution to the data format heterogeneity problem. Based on XML, it uses external metadata (data about data) to allow applications to "plug and play" seamlessly with data sets in heterogeneous formats. An ESML file can be seen as a set of instructions for the application on how to read and understand a data file. There are three key components that make up the ESML interchange technology: the ESML Schema, ESML Description files and the ESML Library (See the schematic depicted in Figure 1).

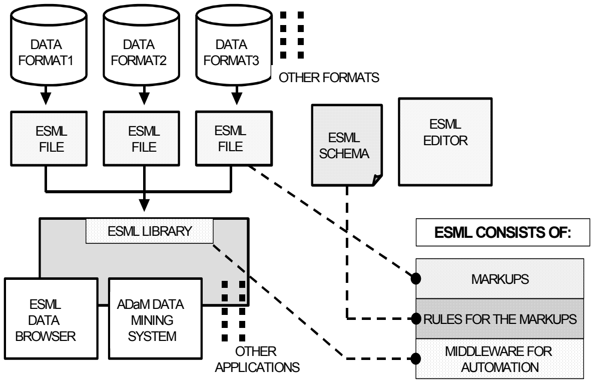


Figure 1: Key Components of ESML

The ESML Schema defines rules that control the bounds for writing ESML Description files. These ESML Description files are the markups written by data producers or consumers to describe particular

*Corresponding author Address:
Rahul Ramachandran
Information Technology and Systems Center
University of Alabama in Huntsville,
Huntsville, AL 35899
email: ramachan@itsc.uah.edu

data sets or files. Because ESML Description files are external files, both data producers and consumers can create and use them at any time. The ESML Library is used by applications to parse the relevant ESML Description file for the structural and semantic information needed to read the data.

2.1 ESML Schema

The ESML Schema defines rules for writing valid ESML Description files. These rules allow users to describe three important aspects of the data file: Content, Structure and Semantics.

a. Content Metadata:

Content metadata describes the scientific properties of a dataset in human-readable terms. Although this metadata may be parsed by the ESML reader, it is not necessarily “understood” by the computer. The content metadata documents the geophysical measurements, origin and pedigree of the data, information that is not necessary for automated data manipulation. The information transcribed by content descriptions is typically used for searching and locating information about data sets. For example, content descriptions might tag a data file using the element <*discipline*> Atmospheric Chemistry</*discipline*>.

b. Syntactic Metadata:

The syntactic metadata are used to describe the details of the data structures within a data file in term of bits, bytes and records. All the syntactic metadata in an ESML Description file are bound by the ESML element tags <*SyntacticMetaData*> and </*SyntacticMetaData*>. The *SyntacticMetaData* element handles different data formats and types (see Figure 2). Currently, the ESML Schema and Library support free formats, such as Binary and ASCII text, the structured format, HDF-EOS as well as other formats such as GRIB and McIDAS. Figure 2 also shows the details of the tags and elements that can be used to describe an ASCII data file.

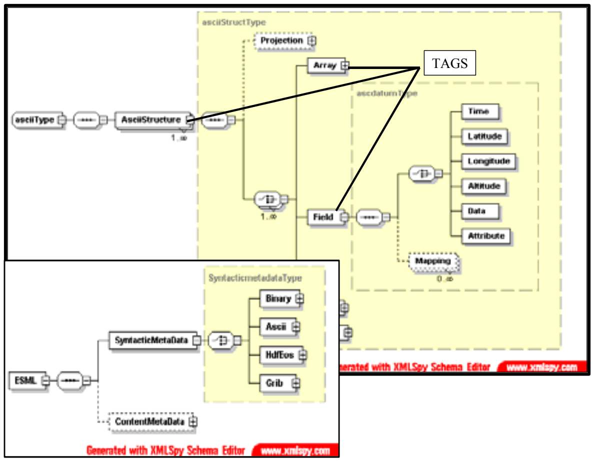


Figure 2: Top level ESML Structural Schema and a detailed view of the ASCII Structure

The ESML Schema has been designed to be extensible such that other data format elements can be added without affecting the existing design. As the project evolves, more and more data formats such as HDF, CDF, netCDF and others will be added to the ESML Schema.

c. Semantic Metadata:

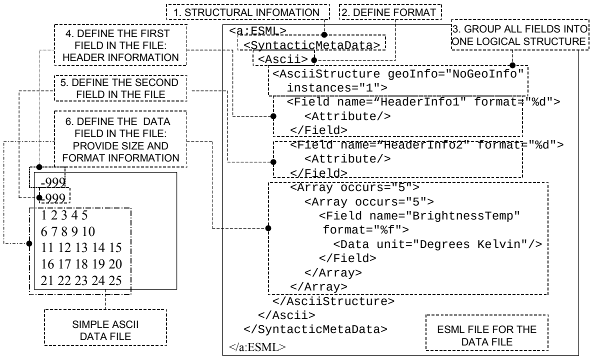
Semantic metadata is used to give meaning to certain elements described by the syntactic metadata. These metadata elements allow the parser to assign meaning to the data elements in terms other than bits and bytes. The semantic metadata is embedded in the syntactic metadata descriptions, within the <*Field*> tags. The basic ESML Schema design goals were not only to allow applications to read the data, but also to be able to spatio-temporally navigate the data and read the data in actual scientific values. Thus, the ESML Schema allows variables in the data file to be tagged semantically as *Attribute*, *Time*, *Latitude*, *Longitude*, *Altitude* or *Data*. Semantic tags can also be used to specify scaling or other preprocessing that may be necessary to convert stored data values to proper geophysical units. For example, in some cases data measurements are stored as integers instead of floating point numbers to save space. If these values are to be used in analysis, proper scaling has to be performed. The ESML Schema provides such means by allowing a user to specify an equation for data conversion using standard C language math notations.

2.2 ESML Description Files

ESML Description files contain the markup description for a collection of data files using the rules defined in the ESML Schema. An example ESML Description file for a very simple ASCII data file is shown in Figure 3. The data file consists of two header fields and a single two dimensional data field. The ESML Description file begins with a specification of the file structure (3.1). This is followed by a description of the data file format, which in this case is ASCII (3.2). The entire file is grouped in a single logical structure (3.3) with no navigation information. The first two *Fields* are then specified with their names and format types (3.4, 3.5). These two *Fields* are subsequently tagged as attributes. The next description is the data field (3.6). The two dimensional array is specified by nesting *Array* elements and setting the array dimensions to *occurs* in the array element. The *Field* is also nested within these array elements, specified with a name and format. This *Field* element is tagged as *Data* with units of the set as “degrees Kelvin”. This semantic tag instructs the parser to read these fields and return the values.

2.3 Library

The ESML Library provides applications with software routines to read and interpret data files based on the ESML descriptions. The ESML Library also provides remote access capabilities to data files via HTTP. The ESML Library is currently available for WINDOWS and LINUX operating systems.


1. STRUCTURAL INFOMATION
2. DEFINE FORMAT
3. GROUP ALL FIELDS INTO ONE LOGICAL STRUCTURE
4. DEFINE THE FIRST FIELD IN THE FILE: HEADER INFORMATION
5. DEFINE THE SECOND FIELD IN THE FILE
6. DEFINE THE DATA FIELD IN THE FILE: PROVIDE SIZE AND FORMAT INFORMATION
SIMPLE ASCII DATA FILE
ESML FILE FOR THE DATA FILE

```
-999
-999
1 2 3 4 5
6 7 8 9 10
11 12 13 14 15
16 17 18 19 20
21 22 23 24 25
```

```
<a:ESML>
  <SyntacticMetaData>
    <Ascii>
      <AsciiStructure geoInfo="NoGeoInfo"
        instances="1">
        <Field name="HeaderInfo1" format="%d">
          <Attribute/>
        </Field>
        <Field name="HeaderInfo2" format="%d">
          <Attribute/>
        </Field>
        <Array occurs="5">
          <Array occurs="5">
            <Field name="BrightnessTemp"
             format="%f">
              <Data unit="Degrees Kelvin"/>
            </Field>
          </Array>
        </Array>
      </AsciiStructure>
    </Ascii>
  </SyntacticMetaData>
</a:ESML>
```

Figure 3: Steps to Write an ESML Description file for an ASCII data file

2.4 Other ESML Products

2.4.1 ESML Editor

The ESML Editor provides scientists with a user-friendly interface to create an ESML Description file without having to deal with the underlying XML tags. The ESML editor also allows users to validate their ESML Description files against the rules specified in the ESML Schema.

2.4.2 ESML Data Browser

The ESML Data Browser is designed to give scientists the ability to view any data file that has an ESML description. Scientists can use this tool to test ESML Description files and to view the data values stored in the files.

3. ESML APPLICATION: MODIS/CERES COLLOCATION

In order for the scientists to exploit scientific datasets in their research, laborious and time consuming preprocessing procedures need to be carried out to customize the data for use in a particular research project. These procedures involve understanding the data format and structure in addition to the actual content. The next step involves writing code to actually read the data, perform preprocessing operations such as subsetting, visualization, etc., and finally either exporting it into their program or more commonly rewriting it into a simpler format. The preprocessing overhead often causes redundancy of effort. Scientists using similar data sets end up writing redundant programs to preprocess the data. ESML will eliminate some of this preprocessing overhead. To illustrate this, an example application of ESML in satellite remote sensing is currently in progress. The science aspect of this application focuses on a multi-sensor approach to examine the radiative effects of aerosols.

It has long been recognized that aerosols play a critical role in the radiation balance of the Earth-atmosphere system. One of the major improvements expected within the lifetime of the ESE mission is to combine different satellite sensors, ground-based and *in situ* measurements to measure and validate the effect of aerosols. Within Terra there are several instruments that can be used to examine the effect of aerosols. The Moderate Resolution Imaging Spectroradiometer (MODIS), a multi channel satellite imager, is used to detect aerosols and provide a global picture of aerosol distribution and thickness. However, there is a need to combine this information with broadband measurements from the Clouds and the Earth's Radiant Energy System (CERES) onboard Terra. Additional information on the angular distribution of aerosols can be obtained from the Multi-angle Imaging SpectroRadiometer (MISR) also onboard Terra.

Typical processing steps include obtaining data, writing code to extract information from each data set by region and by parameter from the HDF-EOS files. This is probably the most time consuming part of the preprocessing where a separate piece of computer code (usually in C) is written to extract information from the HDF-EOS file and output in binary format for later use. After information is extracted from each sensor, collocation is performed to prepare the data for scientific analysis [3]. Coupling the ESML library with the Collocation Algorithm removes the data format complexity (See Figure 4.). There is no need to translate data from one format to another.

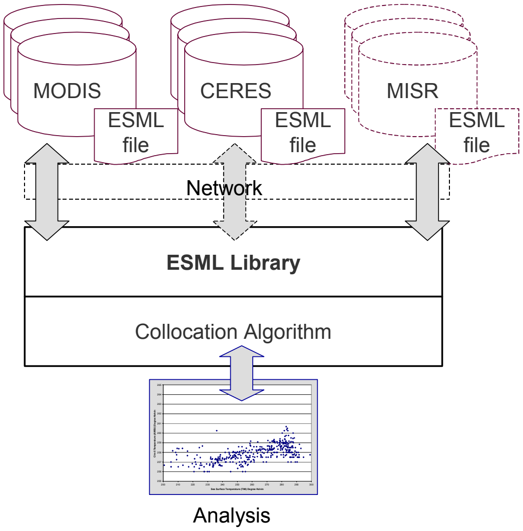


Figure 4: MODIS/CERES Collocation Application

The ESML enabled collocation software can now be extended to work on other kinds of data formats. Scientists can select different fields for collocation by modifying the semantic tags in the

ESML description file. In addition to solving the data format issue, the coupling with the ESML Library also allows the collocation software to access and use remote data files distributed across the network. A sample collocation plot generated by this application is depicted in Figure 5. Work is underway to provide this application to other scientists by converting it into a web service.

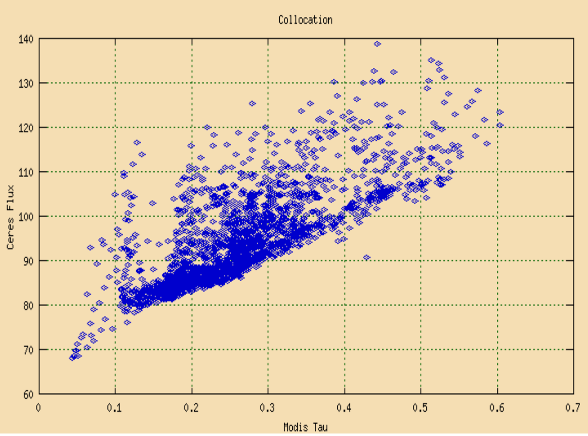


Figure 5: Sample Collocation plot generated by the application

4. SUMMARY

The ESML provides an elegant solution to the data format heterogeneity problem. The combination of the ESML Description files, Schema and the ESML Library-enabled applications form a new interchange technology that will allow applications to utilize different data sets seamlessly. ESML is not a new data format. Rather, ESML allows the scientists to use data in a wide variety of formats and yet still achieve interoperability with different applications. ESML also helps software developers, enabling independently developed applications and services to effectively utilize a wide variety of distributed, heterogeneous data products. An example application that collocates different parameters in MODIS and CERES data files was described in this paper. Additional information about the concepts, tools, and products mentioned in this paper can be obtained at the ESML website [4] (http://esml.itsc.uah.edu).

5. ACKNOWLEDGEMENTS

This research work has been funded by NASA's Earth Science Technology Office. The authors would like to acknowledge Bruce Beaumont, Andrew McDowell, Matt He, Matt Smith, Xiang Li, Sunil Movva, Amit Seth and Jun Wang for their contribution to this research.

6. REFERENCES

[1] R. Ramachandran, H. Conover, S. Graves and K. Moe, 2002. Earth Science Markup Language. Submitted to Computers & Geosciences.

[2] R. Ramachandran, M. Alshayeb, B. Beaumont, H. Conover, S. Graves, X. Li, S. Movva, A. McDowell and M. Smith, 2001. Earth Science Markup Language: A Solution for Generic Access to Heterogeneous Data Sets. Earth Science Technology Conference, Maryland.

[3] Christopher, S. A., and J. Zhang, 2002: Shortwave aerosol radiative forcing from MODIS and CERES observations over the oceans. Geophysical Research Letters, in press, http://vortex.nsstc.uah.edu/~sundar/papers/grl_2002_revise.pdf.

[4] Earth Science Markup Language Web Site: http://esml.itsc.uah.edu.