ARCHIVING OF ATMOSPHERIC DATA: DATA FORMATS AND DATABASE

M. De Mazière and M. Van Roozendael
Belgian Institute for Space Aeronomy, Brussels, B-1180, Belgium

B. R. Bojkov
Norwegian Institute for Air Research, Kjeller, N-2027, Norway

J. de la Noë
Observatoire de Bordeaux, Université Bordeaux 1, Floirac, F-33270, France

E. Mahieu
Institut d'Astrophysique et de Géophysique, Université de Liège, Liège, B-4000, Belgium

R. Neuber
Alfred-Wegener-Institut für Polar und Meeresforschung, Potsdam, D-14473, Germany

ABSTRACT

In the frame of the European project COSE, Compilation of Atmospheric Observations in support of Satellite measurements over Europe, new paths have been explored for data archiving. A number of guidelines have been proposed, including very detailed metadata requirements, for creating well characterized data sets. The resulting generic data submission guidelines have been established to serve a broad community of atmospheric sciences. The metadata in particular will be the key to data quality assurance and quality control, to data indexing and help in the implementation of a relational database with search and query facilities.

1. INTRODUCTION

In the last decade, European Scientists have lend great support to the Network for Detection of Stratospheric Change (NDSC; http://www.ndsc.ncep.noaa.gov/). The European contributions to the official NDSC database at NOAA are mirrored in NILU's Atmospheric Database for Interactive Retrieval (NADIR; http://www.nilu.no/nadir.html), a European atmospheric data centre hosted by the Norwegian Institute for Air Research (NILU). Since NADIR's inception in 1991, the database has grown rapidly, and it appears that the existing flat file structure no longer satisfies the user needs and the database management requirements. Currently, the database structure is linked to the NASA-Ames format (Gaines and Hipskind, 1990), which is used for data submission. Although NASA-Ames is a relatively simple and versatile ASCII file format, it is neither well suited as a submission format for performing an effective quality assurance and quality control (QA/QC), nor for building a relational database with advanced search and query facilities.

In the frame of the European project COSE, Compilation of Atmospheric Observations in support of Satellite measurements over Europe (De Mazière, 1999; http://www.nilu.no/projects/nadir/cose/cose.html), new paths have been explored for data archiving which will have positive implications for the NADIR database and its users in the near future.

Within the COSE project, ground-based network data from 15 partners, operating different types of NDSC monitoring instruments in Europe, are collected with the goal of serving various scientific and satellite users. It focuses on a selection of key data products like O_3, NO_2, and Cl_y total column abundances and/or vertical profiles, and some dynamical tracers (HF, N_2O) column abundances.

It has been agreed within the COSE consortium to adopt NASA's Common Data Format - CDF (Goucher and Mathews, 1995) for data file submission to NADIR. CDF is a binary file format with many inherent advantages: it is well established, mature, platform-independent, and numerous tools exist in standard software packages to handle and visualize CDF file contents (Mathews & Towheed, 1995; Research Systems Inc., 1994). In addition, the CDF file structure is based on the concept of a skeleton, a type of template which defines all the file's variables, including the measurement data, support data and metadata (http://nssdc.gsfc.nasa.gov/cdf/cdf_home.html). This concept and the fact that the CDF file is self-describing facilitate QA/QC.

Guidelines for the metadata have been developed to fully characterize the data sets. For example, the dataset's metadata may include experimental characteristics such as the measurement integration time and its resulting data uncertainties; and or the post-measurement data analysis methodologies, such as algorithm specifications and

IRS 2000: Current Problems in Atmospheric Radiation
Smith and Timofeyev (Eds.), ISBN 0-937194-43-3

references. The adoption of common metadata rules also enables the implementation of a real relational database, with quasi-automatic data cataloguing. Based upon the guidelines, well-defined CDF skeleton files for each experimental technique involved in COSE have been defined (Bojkov et al., 2000). It must be emphasized however that the metadata guidelines are independent of the CDF data format and may be applied to any modern file format. In addition, they have been made general enough to serve a broad community of atmospheric (and other geophysical) data providers.

In this paper the data file submission format and metadata guidelines developed in the COSE project are presented. The implications for the construction of a relational database will be illustrated.

2. METADATA

A well-defined data set must consist of a number of variables, some pointing to the data and their associated attributes, others referring to the necessary data documentation (metadata) and support data. In addition to the metadata describing a particular variable, a number of general attributes must be defined that apply to the whole data set under consideration. Guidelines have been set to harmonize: 1) the attributes nomenclature, and 2) the options for the attribute entries. These will be described hereinafter. A detailed description of the COSE data format requirements (Bojkov et al., 2000) is available.

2.1 General attributes

The general attributes describe, in broad terms, the origin, the ownership, and data type found in a given file. These entries have been divided into 3 sub-categories: 1) attributes referring to the project/campaign the data are related to, 2) attributes dealing with the data Principle Investigator (PI) and Data Originator (DO), and 3) attributes giving general information on the data contents. The general attribute guidelines are summarized in Table I.

Table I. Guidelines concerning the data set general attributes. The 3 categories (project association, data responsibilities, data content; see text) are separated by bold horizontal lines. F: free-formatted, G: according to guidelines, T: to be selected from a table, M: multiple entries are allowed.

ATTRIBUTE		DESCRIPTION	EXAMPLES
DISCIPLINE	M T	field >class >subclass.	"atmospheric chemistry >optical remote sensing > ground based"
PROJECT_ACCESS	M T	Projects associated with the data, inherently used to set the data access rights at the database.	"COSE", "NDSC"
PROJECT_ASSOC	M	All other projects the data are associated to.	
RULES_OF_USE	T	E.g., reference to a data protocol.	
SKELETON_REF	T	The reference skeleton file for the actual data set.	
PI_NAME	G	Complete identification of the PI who is responsible for the data set and/or project.	
PI_AFFILIATION	G		
PI_COORDINATES	F		
PI_EMAIL	G		
DO_NAME	G	Analogously, complete identification of the person who is responsible for the actually submitted data set, the so-called data originator	
DO_AFFILIATION	G		
DO_COORDINATES	F		
DO_EMAIL	G		
ACKNOWLEDGEMENT	F	The data acknowledgement any data user should give.	
DATA_SOURCE_DESC	F	Complete description of platform, instrument, measurement and data type.	"Jungfraujoch hourly level 2 data Bruker 120HR FTIR spectrometer"
DATA_GROUP	T	Type of data > subtype	"experimental > stationary"
DATA_TYPE	T	time resolution // data level Time resolution: s, m, h, d, o ; Data levels: 0, 1, 2 or 3.	"h2"
DATA_SOURCE_CODE	T	Platform_datatype_instrument code.	"issj_h2_ftir2"
DATA_ELEMENTS	M T	Parameter > matrix	"O3 > profile", "H2O > insitu", "emission > limb"
DATA_VERSION	G	Nnn	"001"
DATA_MODS	F	Data modifications history, referring to data version number.	
DATA_CAVEATS	F	Any warnings a user of the data should be aware of.	
FILE_NAME	G	Platform_datatype_instrument_startdate.ver	"issj_h2_ftir2_19990301.001"
FILE_GEN_DATE	G	The generation date of the data file	"19991227"
PARENT	G	File name of parent data	"issj_h1_ftir2_19990301.001"

Table II. Guidelines concerning the variable attributes. Below the bold line, the attributes specify default characteristics for display only. T: to be selected from a table.

ATTRIBUTE		DESCRIPTION	EXAMPLE
VAR_DESC		Complete description of the data the variable points to.	"Ozone total column from Dobson spectrometer "
DEPEND_0 DEPEND_1 DEPEND ...		Specification of the dependencies of each dimension of the variable; dimension 0 corresponds to time. This attribute may point to another variable in the same data set.	
VAR_UNCERTAINTY		points to another variable in the same file specifying this variable associated uncertainty(ies)	
VAR_TYPE		specifies whether "primary", "support" or "metadata"	
VAR_NOTES		Additional notes concerning the variable	
UNITS	T	The units in which the variable data are specified.	"mPa"
SI_CONVERSION		Conversion of the units to corresponding SI units	"10E-3 > Pa"
AVG_TYPE		Averaging technique used.	"rms"
MONOTON		"false", "decrease" or "increase"	
VALIDMIN		Valid data minimum, or detection limit.	
VALIDMAX		Valid data maximum, or saturation limit.	
FILLVAL	T	Missing data fill value.	
FIELDNAM		Concise variable description for heading a tabular data display.	"Ozone"
FORMAT		Display format	"F8.3"
SCALEMIN		Display graph default scale minimum.	
SCALEMAX		Display graph default scale maximum.	
LABLAXIS		Display graph axis label.	"O3 (mPa)"

The entries are either free formatted (F), either the format is subject to given guidelines (G), or the entries are to be chosen from a table of options (T); in some cases multiple entries are allowed (M).

2.2 Variable attributes

The variable attributes (Table II) describe the characteristics of the data represented by the given variable. For some, the entries are to be selected from a table (T) with the goal of harmonizing the data set nomenclature, thus facilitating the incoming data QA/QC and efficient data exploration. A prime example would be the homogenization of the names and spelling of reported chemical species, of the measurement units in which they are specified, and the definition of fill values to be used in case of missing data.

2.3 Tables

Tables have been initiated to consider most of the existing atmospheric modeling and experimental scenarios. These option tables describe numerous types of experimental platforms, measurement or modeling techniques, observed / modeled species, and data types and units, and will be updated as future needs change or evolve. Table III, for example, specifies the atmospheric disciplines options defined in COSE/NADIR to describe the field the submitted data relate to.

Table III. Entry options for the global attribute 'DISCIPLINE'. Any combination of field, class, and subclass is allowed.

FIELD	CLASS	SUBCLASS
Atmospheric chemistry	Optical remote sensing	Ground based
Atmospheric dynamics	Radar remote sensing	Space borne
Atmosphere *(Chemistry & dynamics)*	In situ sampling	Balloon borne
Stratospheric chemistry *(Purely stratospheric)*	Model	Aircraft
Stratospheric dynamics *(Purely stratospheric)*		Rocket
Stratosphere *(Chemistry & dynamics)*		1D box
Tropospheric chemistry *(Purely tropospheric)*		2D
Tropospheric dynamics *(Purely tropospheric)*		3D
Troposphere *(Chemistry & dynamics)*		4D VAR assimilation
PBL chemistry *(Purely boundary layer)*		4D sequential assimilation
PBL dynamics *(Purely boundary layer)*		GCM
PBL *(Chemistry & dynamics)*		Lagrangian trajectory
		Eulerian trajectory

3. DATABASE IMPLEMENTATION

The harmonization of the data set structure enables the development of data QA/QC, data indexing, and therefore the implementation of a relational database with search and query facilities. It is can be clearly seen in Figure 1 that the file content control (QA/QC) and data base indexing process hinges on the homogeneity of the submitted file's metadata through the use of the CDF skeleton. For any measurement technique involved in COSE, a typical reference CDF skeleton is being defined. It contains all essential attributes to be reported and an example of the main relevant variables. It sets the structure of any CDF file that may be submitted in this field. Any data provider will build his own skeleton in compliance with the relevant reference one. If a deviation from the reference is desirable, this will have to be signaled to the database management and will be subject to approval by the database and project coordinators.

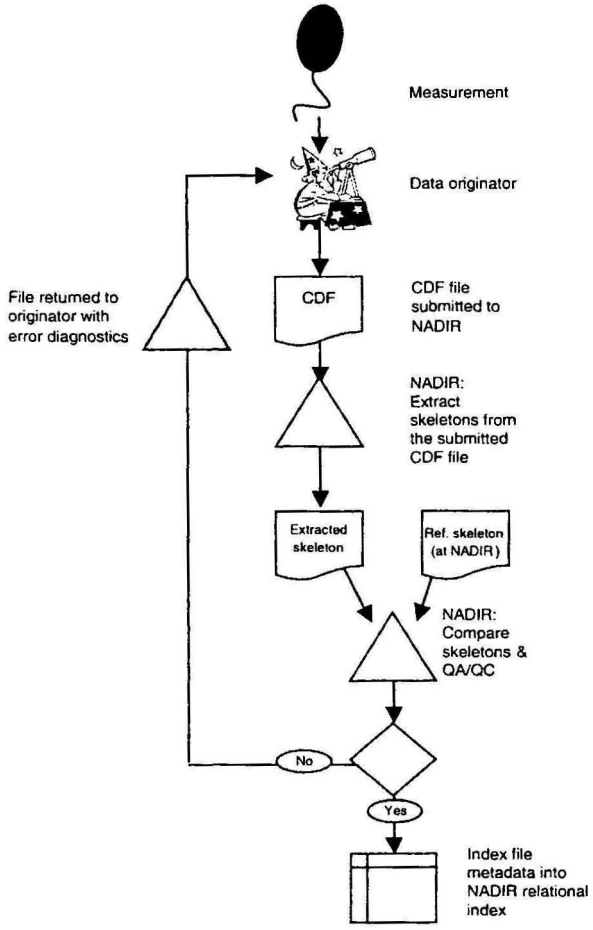

Fig. 1. A schematic diagram of the quality control and data indexing for a CDF file submitted to NADIR.

ACKNOWLEDGMENTS

This work was financially supported by the European Commission DG XII via the COSE project (contract ENV4-CT98-0750). The help and valuable comments of B. Quaghebeur and D. Heynderickx, BIRA-IASB, and T. Krognes, NILU, are greatly appreciated. Thanks are due to O. Lezeaux, IAP/Univ. Bern, and C. David and S. Godin, SA/CNRS, for revision of the CDF skeleton files.

REFERENCES

Bojkov, B.R., M. De Mazière, T. Krognes, 2000, Metadata guidelines for atmospheric sciences: a prototype CDF format from the EC project COSE, NILU Technical Report, in press

De Mazière, M., COSE: Compilation of Atmospheric Observations in support of Satellite measurements over Europe, 1999, ESAMS '99, European Symposium on Atmospheric Measurements from Space, WPP-161, pp. 347-353, European Space Agency.

Gaines, S.E., and Hipskind R.S., 1990, Format Specifications for Data Exchange, NASA/Ames Research Centre, California, U.S.A.

Goucher, G. W., and G. J. Mathews, 1994, A Comprehensive Look at CDF, NSSDC/WDC-A-R&S 94-07, NASA/Goddard Space Flight Center, Maryland, U.S.A.

Mathews, G. J., and S. S. Towheed, "OMNIWeb: The First Space Physics Data WWW-Based Data Browsing and Retrieval System," Computer Networks and ISDN Systems, Proceedings of the Third International WWW Conference, Vol. 27, No. 6, April 1995, pp. 801-808.

Research Systems Incorporated , 1994, IDL Scientific Data Formats, Version 3.6, Boulder, Colorado, U.S.A.