# DESIGN OF DATA SERVER FOR CEOP DATA

TOSHIHIRO NEMOTO
*Institute of Industrial Science, University of Tokyo*
*4-6-1, Komaba, Meguro-ku, Tokyo, Japan, 153-8505*

EIJI IKOMA
*Center for Spatial Information Science, University of Tokyo*
*4-6-1, Komaba, Meguro-ku, Tokyo, Japan, 153-8505*

MASARU KITSUREGAWA
*Institute of Industrial Science, University of Tokyo*
*4-6-1, Komaba, Meguro-ku, Tokyo, Japan, 153-8505*

On the CEOP (Coordinated Enhanced Observing Period) project, in order to improve our understanding of water and energy and fluxes and reservoirs over land areas, large amount of data are being collected and archived. In this paper, we explain the design of the data archiving and analysis server for the CEOP data at Institute of Industrial Science, University of Tokyo, which we are now constructing to improve availability and usability of the CEOP data. The preliminary data server has been already implemented and it is used for experimental operation. This paper also explains their usage in addition to the implementation of them.

## INTRODUCTION

Water and energy cycle are of fundamental importance to the climate system, but they are not handled well enough within climate and numerical weather prediction models. On the CEOP (Coordinated Enhanced Observing Period) project, in order to improve our understanding of water and energy and fluxes and reservoirs over land areas, large amount of data are being collected and archived. The data consist of three kinds of data, that is in-situ data, satellite data and model output data. The in-situ data are a temporal series of air temperature, pressure, humidity, precipitation and so on at 36 reference sites around the world. The satellite data are remotely sensed data from the operational satellites, such as TERRA, AQUA, TRMM, NOAA and so on. The model output data are generated by numerical weather prediction centers. These data have various dimensions, spatial and temporal resolutions, precision, formats, coordinate systems. The total amount of the data is almost 100TB per year.

In this paper, we explain the design of the data archiving and analysis server for the CEOP data at the University of Tokyo, which we are now constructing to improve availability and usability of the CEOP data. The server manages all of the data including meta data together. It uses tape library system and disk arrays to store them, however, the location of data is hidden from users. The users can retrieve data without considering data

location. The server provides the users with menu-based integrated graphical user interface for data retrieve and analysis. The connection between user's clients and the server is based on Web system. The users can access all kinds of data through the same interface without taking account of data type. The users can view the retrieved data as graphic charts or bitmap images depend on their dimension directly from the server. Some analysis operations such as average, difference, correlation, and so on can be applied into one or more retrieved data on the server through the GUI.

The preliminary data server has been already implemented and it is used for experimental operation. This paper also explains their usage in addition to the implementation of them.

## DATA TO BE ARCHIVED

There kinds of data are planned to be archive on the CEOP project. They are in-situ data, model output data and satellite data.

### In-situ Data

The in-situ data are a temporal series of the values observed at 36 reference sites around the world. Each reference site has one or more stations. The in-situ data are divided into three categories, namely surface observation, subsurface observation and upper air observation. The surface observation consists of air temperature, pressure, humidity, precipitation, heat flux, radiation and so on at the ground level. The subsurface observation is composed of soil temperature, soil water content and soil heat flux for 2cm to 175cm depth. The upper air observation consists of air temperature, humidity, pressure and so on measured by radiosonde. The all values are not always observed at a reference site. The sorts of the observed values and observation frequency depend on the reference site. The total amount of in-situ data for two year and three months is almost 600MB.

### Model Output Data

The model output data is the gridded values from global forecast model or assimilation system generated by 10 numerical weather forecast centers. Two types of model output, namely gridded data and site-specific time series at each of the reference site are planned to archive. The latter time series are designated as MOLTS (Model Output Location Time Series). The gridded data are three dimensional data and each cell has several prognostic variables such as air temperature, humidity, pressure and so on. The forecast length, assimilation intervals, the grid systems of the models and also the variables in models are different each other. The MOLTS data are one dimensional time series of variables at the reference site extracted from gridded data. The total amount of model output data is almost 20TB.

**Satellite Data**

The satellite data are remotely sensed data from the sensors on the operational satellites such as DMSP SSM/I, TRMM TMI, TRMM PR, GMS S-VISSR, NOAA AVHRR, TERRA/AQUA MODIS, AQUA AMSR-E and so on. The satellite data are two dimensional data and each sensor has one or more channels. Though these data are geometrically corrected by the data supplier, their resolutions depend on the sensor and the channel. The observation time may vary day by day even if the same satellite. The total amount of satellite data is almost 200TB.

## SYSTEM DESIGN

**Design Policy**

The key concept of our data server system is easy use. To integrate in-situ data, model output data and satellite data the scientists have to do a lot of bothersome operations before they actually analyze them. First they search and retrieve appropriate data. After they obtain them, they have to convert a file format to be suited to the tool they use. Then they may reproject, regrid and subsample the data to make them same projection, same coordinate and same area of interest. They may also interpolate the data to make them same resolution. In addition to the spatial arrangement, they may rectify the data to synchronize the temporal resolution. After these operations, the scientists can compare the different types of data. This preprocessing is not avoidable, but it is a laborious work. Especially, the scientists concerning global environmental data are not always skillful at handling a large amount of data. Consequently, it is expected that the scientists can access and analyze data easily without special knowledge about database and computer systems. In addition, the users for CEOP data are scattered around the world. Their computer environments vary. To avoid using special software and hardware is also an important concept. It is necessary to support as many users as possible in order to improve the usefulness of data.

**System Architecture**

The preliminary system is based on a client-server model. The architecture of the system is shown in Figure 1. The communication protocol between server and client is HTTP. The HTTP is not always suit for data transfer, however, it is widely used and therefore it may cause few troubles especially about a fire wall. Accordingly we adopt HTTP as the communication protocol. The requests from clients are at first received at the HTTP server and then they are sent to the data manager. The data manager is a servlet program. It receives the requests from clients through the HTTP server and then it generates SQL commands for data search or executes analysis operations according to the user requests. One dimensional data such as in-situ data and MOLTS data are stored in DBMS, however, two and three dimensional data such as gridded model output data and satellite data are stored as files on the hierarchical file system and only their metadata are stored in DBMS. There are several reasons why we do not manage two or three dimensional

data as Large Objects (LOB) in DBMS. First, the accessing a LOB in DBMS is slower than accessing a file[1]. Second, existing implementations of LOBs tend to lack support for the hierarchical storage management system. Although the gridded model output data and the satellite data are stored on the hierarchical file system, small images around the reference sites clipped from the global data are stored on disks. The global data are too large to store them on disks and most of the users do not need global data. Generally the scientists need only the values around the reference sites to compare the values from ground observations and that of model output or that of remotely sensed data. To store the small portions on the disks instead of hierarchical file system we can reduce the response time. The location of the data is hidden from the users. The users do not have to consider where the data are stored. The data server automatically migrates and retrieves the appropriate data from DBMS, disks or hierarchical file system as the user requests and sends them to the clients. The DBMS manages the one dimensional data and the metadata for all data. We use a commercial DBMS and JDBC for the connection between DBMS and the data manager.

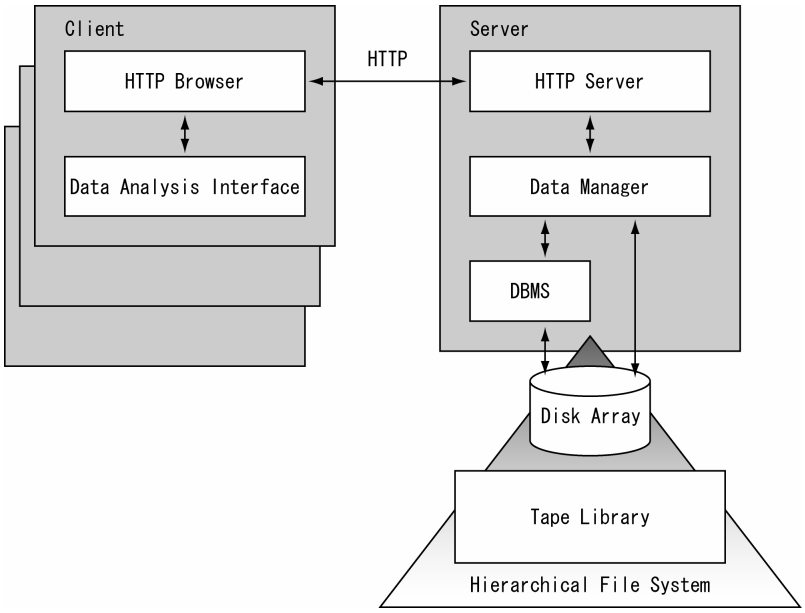


Figure 1. System architecture of data server at Institute of Industrial Science, University of Tokyo

The graphical user interface in the client system roughly consists of two parts, HTTP browser and data analysis interface. The communication between clients and servers are based on HTTP and the data analysis interface is written in JAVA. Accordingly, the client does not need any special hardware or software. Only HTTP browser and JAVA

runtime environment are required. Since many kinds of current computers and operating systems support HTTP browser and JAVA runtime environment, the GUI for the data server works on many kinds of computers. The usage of the data analysis interface is described in the next section.

## USAGE OF DATA ANALYSIS INTERFACE

The user accesses to the data server page through the HTTP browser at first and then authentication page is shown. After the user passes the password check, the user can access to the archived data. The requested data is specified by three items, namely reference site name, data name and temporal period. The available reference site and data name are listed in the menus and the user selects one of them (Figure 2). The temporal period is specified by start date and time and those of end. Clicking the button to retrieve, the request is transferred to the server. The server parses the requests, generates the SQL command, sends it to the DBMS and stores the result into the user's area. The results are listed in the data analysis interface window and they are regarded as targets of analysis operations by the data analysis interface. The retrieved one dimensional data can be displayed as a line graph and also as an ASCII text. The retrieved two dimensional data can be displayed as a series of bitmap images (Figure 3). A temporal series of the selected two dimensional data also can be animated (Figure 4). The animation viewer supports not only simple animation but also supports watching frames forward and backward step by step. In addition, in the animation viewer window the user can expand or reduce the images.

The analysis operation is executed to the retrieved data. The user selects one or more retrieved data and then pushes the appropriate analysis command button. The analysis request is send to the data server and the data manager executes the operation to the selected data in the user's area. In the current preliminary version, only simple analysis commands such as average, difference and so on are available. The data processed by the analysis operation are also stored into the user's area and they are listed in the data analysis interface window. They can be targets of analysis operations. The user can apply the analysis operations onto the processed data again and again. More than two data can be bundled and can be displayed in a single graphic chart (Figure 5). The analysis interface can also draw a scatter diagram. It can extract a temporal series of values at the reference site from a temporal series of two dimensional data. The extracted values can be handled as if it was a one dimensional data. To bundle these data the user can compare two or more type of data. For example, bundling in-situ data and the extracted values from the satellite images, their values are drawn in a graphic chart and the user can compare them.

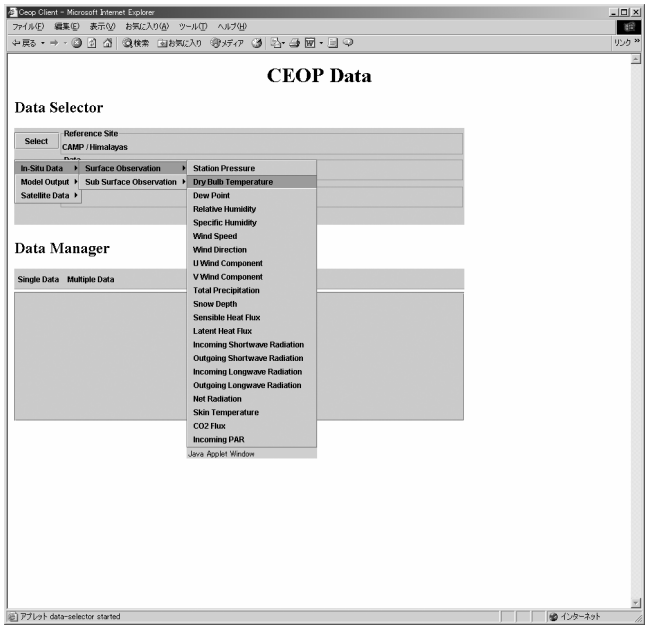


Figure 2. Menu-based data selection

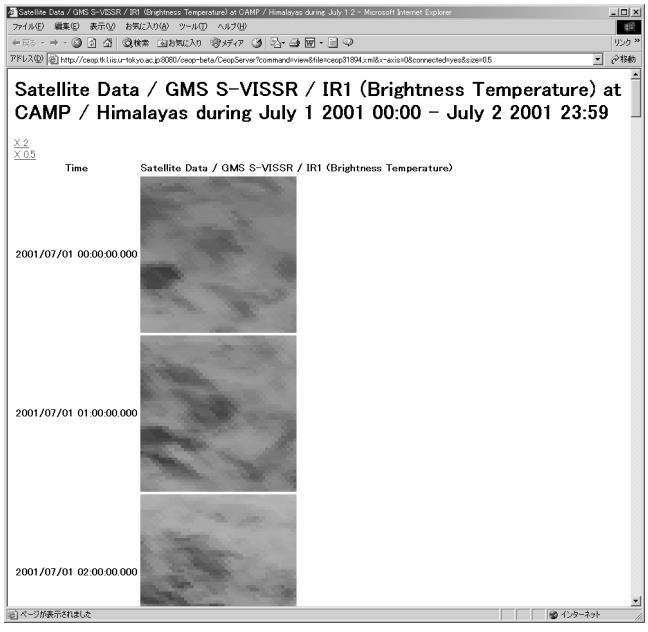


Figure 3. Image viewer for satellite data

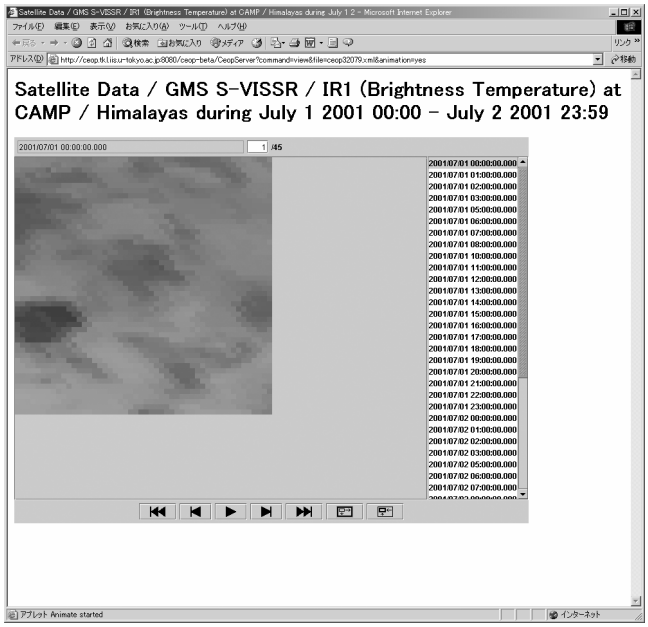


Figure 4. Animation tool for a time series of satellite images

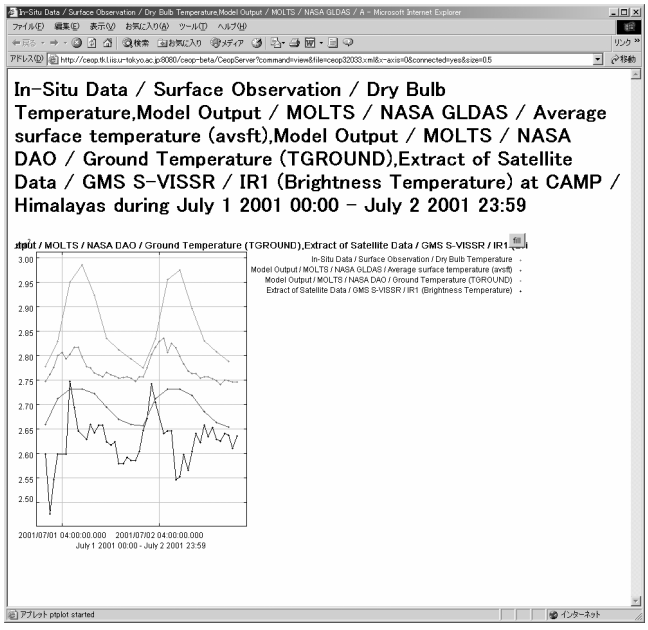


Figure 5. Graphic chart for multiple one dimensional data

## CONCLUSIONS

This paper outlines the data server for CEOP data at Institute of Industrial Science, University of Tokyo. The main object of the system is to make it easy to access and analyze the data. The scientists concerning water cycle should process large amount of data, however, they are not always skillful at handling large data. Though our current preliminary data server supports only simple operations, it may be helpful to them. We are now implementing more analysis functions and improving the system. The data server will be open to the CEOP community tentatively in the near future.

## REFERENCES

[1] Stolte E., Praun C., Alonso G. and Gross T., “Scientific Data Repositories – Designing for a Moving Target”, ACM SIGMOD 2003.
[2] CEOP Home Page, http://monsoon.t.u-tokyo.ac.jp/ceop/.