Date : 21/05/2007

An Analysis of the data archiving practices of selected scientific agencies in the Philippines

Carina C. Samaniego
Manila Observatory
Philippines

Meeting: **140 Asia and Oceania**
Simultaneous Interpretation: No

WORLD LIBRARY AND INFORMATION CONGRESS: 73RD IFLA GENERAL CONFERENCE AND COUNCIL
19-23 August 2007, Durban, South Africa
http://www.ifla.org/iv/ifla73/index.htm

Abstract

The field of science is changing and the volume of scientific data, in different formats, is growing exponentially. This study will focus on the data archiving practices of three (3) scientific agencies in the Philippines. As a preliminary study, a specific type of data from each agency is chosen to illustrate how it is archived: from data generation up to preservation and to be able to describe the major issues in the archiving of scientific data in the Philippines as seen through these agencies.

Introduction

The International Council for Science (ICSU) defines *data* as digital observations, scientific monitoring, data from sensors, metadata, model output and scenarios, qualitative or observed behavioral data, visualizations, and statistical data collected for administrative or commercial purposes. These can be acquired through the traditional or *in situ* methods (physical sampling) or through the use of modern technologies such as satellites, radars and automatic data loggers. The advancement of technology leads to the increased capacity to collect more data at a wider range and of various types.

Data plays an important role in the development of scientific knowledge. It may confirm or refute an existing theory; or may create a new discovery that would lead to a new and deeper study and requires new set of data to support it. Observed data on the earth and physical sciences, such as astronomy, planetary science, meteorology, etc., provide a baseline for determining rates of change and for computing the frequency of occurrence of natural events, such as El Niño, volcanic activities, earthquakes, typhoons, etc. These are not only useful in predicting the future but also in creating a map of the past and its relation to the future.

There are many problems commonly encountered by agencies dealing with scientific data. One is data explosion. Research on climate and atmospheric science, for example, requires long sets of records for comparative purposes. Thus, there is an

increasing demand for historical climatic data. Another problem is the lack of directories or databases that describe the availability and accessibility of data available elsewhere. Many data are also at risk due to the absence of backups or the medium they are being stored is deteriorating, corrupted or superseded by new technology (NRC, 1995). And there is always the threat of disasters, both man-made and environmental in nature, such as fire, flooding, and theft.

Thus it is important that each agency working in the field of science will develop an effective archiving system for the preservation and further dissemination of its data and endeavors. Data archiving in science should not just be a system of procedures for long-term preservation and access but also places the same emphasis on content and meaning of the collection.

Methodology

The Philippines, as an archipelago, is highly prone to natural disasters. Based on the study of the Centre for Research on the Epidemiology of Disasters (CRED), the Philippines ranked 3rd worldwide with the most number of natural disasters for 2006. Despite of this fact, its geographic location also plays an important role in the study of the influence upon the earth of natural phenomenon such as earthquakes, tsunamis, and typhoons.

The three agencies that are part of this study are the forerunners in providing relevant data and studies in the fields of natural disasters in the Philippines. The Philippine Atmospheric, Geophysical, Astronomical Service Administration (PAGASA) is the official agency in the field of meteorology, operational hydrology, climatology, astronomy and other allied sciences. The Philippine Institute of Volcanology and Seismology (PHIVOLCS) monitors the occurrence of volcanic eruptions and earthquakes and other geotectonic activities. The Manila Observatory (MO), the first official weather and seismological bureau in the Philippines, is a private research institution that deals with projects on climate studies, disaster risk management, air quality monitoring, socio-environmental geomatics, instrumentation and technology development, and solid-earth dynamics.

The types of data studied from each of the agencies are as follows: Meteorological data from PAGASA's Climate and Agrometeorology Branch; seismic data from Phivolcs' Seismological Observation and Earthquake Prediction Division; and MO's meteorological and seismic data. These types of data were selected due to the fact that each agency has an enormous data collection with a decentralized system of management. This is also to illustrate the differences and/or similarities in the archiving of these particular data as practiced by these institutions.

The principle of archiving includes the processes of Appraisal & Acquisition; Arrangement & Description: Reference & Access; Education & Outreach; and Preservation and Conservation. Regardless of type or format, any data or material being archived goes through these basic processes; the difference is on the manner how they are acquired, recorded, arranged, and conserved. Based on these assumptions, a simple framework is followed in this study:

Figure I. Schematic framework of the study

As shown by the diagram, this study will look into the structure of the data archiving system of each institution: starting from the generation of data from **Data Source/s**; appraisal, description, arrangement and preservation of data in the **Archives (or Data Center, or whatever is used as a depository)**; and data access, retrieval and use of the **Designated Community**; and how the system is established, maintained and improved by its **Management**.

PAGASA—Climate and Agrometeorology Branch (CAB)

CAB is the branch of PAGASA that is responsible in the monitoring and collection of meteorological, climatological and agrometeorological data. It operates the Climate Data Section (CDS), which is considered as the Climate Data Center of the Philippines, where all records of collected data are processed, stored and preserved. The specific types of raw data that CAB collects are Synoptic data which are meteorological parameters such as rainfall, temperature, wind direction, etc.; Agromet observation are meteorological and biological data useful in the field of agriculture such as temperature, vapor pressure, cloudiness; and Climatological data which is the measurement of rainfall/precipitation. These data are sourced from PAGASA's monitoring stations located in different parts of the country. Each type of data differs on the frequency in recording/acquisition. Standard forms for each type of data are used to record readings and copies are regularly sent to PAGASA's main office for processing, recording, analysis and storage at the CDS. The raw data are used to produce the following outputs: publications such as the daily and monthly climatic, agroclimatic, and climat data; climate data products such as means/extremes on daily, 5, 7, 10, and monthly intervals, normals and averages; maps showing the climate observations per station which includes cumulative rainfall, percent normal, total rainfall, etc.; station profile; maximum and minimum temperature for a certain month compared with the normal and historical climatic records (attachment to Weather situation in the Philippines); and monthly rainfall statistics.

It currently uses CLICOM which is a standard data management system for WMO member countries. This program enables the management and analysis of data such as the recording, query, statistical analysis, modeling and visualization of climate phenomenon in a given region/area. Vistualization data are interpreted through maps and

charts and reports can be produced such as standard monthly climate publications. It can also validate recorded weather observations and historical climate data.

Researchers do not have direct access to the data. A fee is required for each request depending on the type of processed data, analysis needed or volume of data needed. A letter of request should be sent to the Director or requests can be filled at the CDS or online. Some relevant data and studies are available online at CAB's webpage and a list of available data products is also accessible. Its clientele ranges from the general public to specialized sectors of the research community.

Back up for digital data are stored in the server and CDs. There is another data storage facility inside the compound of PAGASA. The data forms that are received from the monitoring stations are compiled and arranged on shelves at CDS. These are arranged per type of forms, then alphabetically by location and chronologically by date. Digitization is also being done for old data sheets.

PHIVOLCS—Seismological Observation and Earthquake Prediction Division (SOEPD)

The Seismological Observations and Earthquake Prediction Division (SOEPD) of PHIVOLCS monitors and documents the occurrence of earthquakes and their geologic phenomena. It generates data to be used in the mitigation of the hazards brought about by earthquakes. The primary data that PHIVOLCS collects is seismic data. The seismic data are sourced from the 69 stations in its seismic network that monitors and collects data 24-hours everyday. PHIVOLCS uses short period instruments which can accurately locate smaller earthquakes for as low as magnitude 2.5.

PHIVOLCS initially used analog seismographs. The readings or seismograms are directly printed on paper. A copy of the seismograms is regularly sent to the main office and a copy is maintained in each field office. Since 2004, it shifted from analog to digital seismographs which can produce seismograms in printed and digital format. Only the digital seismograms, stored in CDs, are now sent to the main office. Aside from the seismograms, there are no other sources to record seismic data. Data transmission is standard for every field station. Time of transmission of phase readings is 8-10 am on a daily basis via fax, phone or radio.

SOEPD maintains data both in printed digital formats. The seismic and waveform data are used to produce processed data products such as seismic phase catalogue, earthquake catalog, earthquake bulletin, tsunami bulletin, and seismic swarm updates and reports.

This division is tasked to compile a series of earthquake catalogues. The purpose of the catalogue is to provide a complete earthquake data and information on earthquake/seismic studies, risk studies and inquiries from the public in each given year.

All requests for processed data should be coursed through the Office of the Director. Requests can be answered within 7-14 days depending on the type of data and analysis requested. Most requested data are seismicity map, earthquake listing, certification for the occurrence of an earthquake, near an area of concern, hazard maps (groundshaking, tsunami, etc), strong motion data (accelerographs), broadband

seismograph data, and microtremor data. Aside from requests by students and the general public, the above data sets are requested by private consultants for evaluating development projects such as construction of buildings, dams and mining sites and for insurance claims, all of which are for a fee.

The current issues that SOEPD faces regarding data management are data format incompatibility, lack of manpower and programmers to manage the growing number of its data. The move from using analog to digital seismographs requires a system of data migration from one medium to another to make the data accessible. There is also a need for addition programmers that can develop programs that would integrate all data in different formats into one single database. The agency is also restricted from hiring additional personnel or going full automation due to economic reasons.

PHIVOLCS plans to maintain three (3) means of storing its data. The Main office will serve as the main repository of all data. A data library is planned to be set-up inside its Data Receiving Center, designed with controlled temperature and humidity and limited access. The Tagaytay office, a city south of Manila, acts as a mirror site for the main office and is being renovated to establish a data library for both and analog and digital data collection. Lastly, each seismic station is required to keep its own data archive.

Manila Observatory

The Manila Observatory (MO) is currently engaged in various research programs and activities. One of its current projects is the Solid Earth Dynamics (SED) project that handles research work on seismology. The only remaining seismic station of the MO is located in Davao City, which is in the southern most part of the Philippines. It uses analog seismograph that operates 24-hours everyday. This seismic station receives technical support from the US Geological Service as part of its Global Monitoring Network project.

The sole copy of the seismogram generated is sent every two weeks to the Manila Observatory for analysis, reporting and compilation. A report of significant seismic events is submitted and included in the International Seismological Center Bulletin. The seismograms are compiled and stored in the seismology vault located in the grounds of the Observatory. This vault was built during the 1960's together with the solar building when the Observatory transferred to its current location. The vault contains the old instruments used in the study of earthquakes and all the seismograms collected which dates back from 1956 to present. The seismograms are grouped per month on every given year. These are placed one over the other on shelves. The room where the seismograms are stored is located in the far end of the vault. There is no built in ventilation such as air conditioners, and there are insufficient light inside the room.

There are plans to digitize the seismograms, which is still under the consideration of funding agencies where the proposals are presented. Manila Observatory projects and research programs are mostly funded by different local and international agencies.

The Manila Observatory also operates its own digital meteorological station since 2001. It currently uses the Davis Vantage Pro2 Weather Station which records 15 meteorological parameters which are as follows:

Name	Unit of measurement	Frequency (Sampling cycle)
Wind direction	^{o} (Degrees)	5 s
Indoor relative humidity	% RH	10 s
Outdoor relative humidity	% RH	10 s
Instantaneous wind speed	m/s	5 s
Indoor temperature	^{o} C	10 s
Outdoor temperature	^{o} C	10 s
Cumulative rainfall	mm	Varies depending on rainfall frequency
Atmospheric pressure	mbar	15 mins.
Dew point temperature	^{o} C	10 s
Wind chill temperature	^{o} C	5 s
Indoor heat index	^{o} C	Depending on record schedule
Highest instantaneous wind speed in the last 20 samples	m/s	Depending on record schedule
Rate of rainfall per hour (Current cumulative rain – cumulative rain of the previous hour)	mm/hr	Varies depending on rainfall on rainfall frequency
Daily rain	mm	24 hours

Table I. List of data collected by MO's meteorlogical station

The data is logged and saved in comma separated half-year text files. The text files are saved in the computer hard disk and CDs and no print outs are produced. There are no particular processed data outputs being produced using these monitored data. These raw data can be requested for free and these are mostly requested by students working on special projects.

Common issues and challenges

Based on the presented information we can see the similarities on the strategies, issues and challenges that these agencies face in the archiving of their data.

The type of data that these agencies collect is called observational data which is the category of data for earth and space sciences. This type of data is naturally occurring and can never be repeated exactly, thus it should be continuously monitored which results to its increased volume over time. It is necessary to maintain a complete and continuous set of the observed data to be able to arrive at conclusions and to generate processed data sets/outputs. Each of the mentioned agencies does not discard data and tries to maintain all of its data.

Maintaining the stations/sources of data is as important as maintaining the data itself. These agencies operate monitoring stations, most of which are located in different areas around the country. There is always the threat of loss due to theft, flooding, electricity failure which will affect the transfer of data, etc.

The problem is not really of what these agencies have, but more on the system (or the lack) of documentation, access and long-term preservation of the data. Unlike in libraries where you can use the Library Congress cataloging rule to code and arrange the

books, these scientific, numerical, raw data, with their unique parameters and degree of occurrence, makes it more difficult to set and implement a general standard that would cover each of the monitored data. Each of these agencies has its own internal system for the description and arrangement of its data, whether it is in raw or processed form. But each of them also admitted that there is a great need to improve their database systems to integrate the growing volume of the data in different formats. Clearly, it is not acceptable to just always describe that their collection is enormous or to provide an estimation base on years of coverage alone.

Directories and databases available for the users are essential to know what data is available and where it can be found. It is also a good indication of the capacity of these agencies as data and information providers based on the available data. These three agencies utilize their websites to disseminate information about their data collection. Among the three, PAGASA has the most intensive listing of its data products available through the CAB's webpage. This will also be helpful when creating network and collection consortiums in the future to avoid any duplication or data redundancy.

The limited accessibility of users, especially to the raw data, is essential in protecting the data content and to avoid unofficial generation of processed information. The fees required by PAGASA and PHIVOLCS are government directive to support the operations and as an austerity measure. MO, due to its non-profit nature, disseminates its meteorological data free of charge as part of its services. It plans to continue uploading the meteorological data into the website, in real time as it is monitored.

As mentioned, the continuous advancement of the instruments being used affects the data collection system of these agencies. This not only affects the rate of acquisition of data, but also the format of the data it can produce. If preserving data on paper is already a concern, digital data is another consideration. In the case of PHIVOLCS, which uses digital seismographs, it now has to manage both paper and digital seismograms and to design a system of codes and protocols for the accomplishment of regular inventory and integration of metadata and data into their databases. And each format has its own set of preservation requirements due to their physical properties.

Creating back-ups is an important measure in ensuring that there will be available copies of the data for contingency and preservation measures. This does not only involves duplication of the data, but ensuring that these are in usable format over time and that it corresponds to the software or system that the agency uses to access its data. Main consideration should be placed on previously recorded data which might be in mediums that are no longer accessible and in sensitive physical condition. PHIVOLCS' plan to have mirror sites of its data center is a good practice to ensure the safety of its data. Preservation programs always involve large capital investments, both in personnel and technological aspects, and should be planned and executed systematically.

The nature, composition, and roles of these agencies affect how data are being managed. All three agencies have a decentralized system of data management. These agencies are made up of a number of divisions and programs with specific roles and activities. This kind of decentralized system can be an advantage for these agencies. Each division or program has the expertise or knowledge over the type of data it collects and works on. It can set the parameters for the data collection, description, access and use. It

can be a disadvantage over time if there is already considerable increase in the volume of the data and priorities may be divided between improving the acquisition of new set of data and preserving the previously collected data.

A centralized system/archive for all data collection can be beneficial for large data collection. This will ensure that there are personnel with specialization on archiving and institutional programs and procedures that will focus on the entire management of the data—from recording, arrangement, dissemination and preservation.

Conclusion and recommendations

The basic principle of archiving is to ensure that the data, in whatever form or format, will remain stable and accessible over time. The archiving program of institutions is based on their nature, roles and priorities. Scientific agencies, with its important roles as official sources of scientific data and information, should have an established system of data management and archiving. An effective archiving system should not be limited to one area or activity alone; it is an encompassing process that is made up of interlinked activities that support one another. If there is a move to improve data collection, consider also how these can be recorded, how it can be accessed, and made available over time.

These Philippine agencies face the same issues and challenges any institution in any part of the world encounters in terms data archiving: natural threats, financial constraints, lack or limited archiving programs and systems, differences in priorities and management issues and limitations. And admittedly, there is still the general issue of lack or low importance on data archiving among scientific institutions. There is always the question how much to keep, for how long, and in what manner. Given these considerations, it might be difficult at this level to impose data archiving standards for these types of agencies to follow, more so, to plan for a national depository for these types of data. Start from creating an integrated program within each agency: considering their roles, activities, collection, and capacity to manage a data archiving system. An effective archiving program does not lie on the advancement of the software or hardware being used: it's knowing what you have and providing an effective means to access it, not only to the present but for the future as well.

References:

Disasters and climate change-do the math

< http://www.alertnet.org/db/blogs/1564/2007/00/29-163942-1.htm>

Gracy, David B. 1981. An Introduction to archives and manuscripts. Special Libraries Association : New York.

International Council for Science. 2004. Scientific Data and Information. Report of the CSPR Assessment Panel.

National Research Council (NRC). 1995. Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information

Resources. Commission on Physical Sciences, Mathematics, and Applications. National Academy Press : Washington D.C.

Acknowledgments:

The Manila Observatory, Philippine Atmospheric, Geophysical, Astronomical Service Administration (PAGASA), Philippine Institute of Volcanology and Seismology (PHIVOLCS), and Swedish International Development Agency (SIDA).

Author's short biography:

Carina C. Samaniego is the librarian of the Manila Observatory since 2001. She earned her Bachelor's degree in Library and Information Science from the University of the Philippines in 2001. She's finishing her graduate degree in Library and Information Science, and archiving as her area of specialization from the same university. The paper she's presenting is part of her graduate thesis. She's currently heading an archiving project under the Ford Conservation and Environmental grant, entitled *Preserving nature's worst: historical accounts on natural disasters in the Philippines.*