

# A Data Model for Climate Data Curation

Soraya Abad Mota, Lelys Bravo de Guenni

Universidad Simón Bolívar. Caracas, Venezuela

V International Symposium on Digital Libraries  
October 2009

# Outline

- 1 Motivation
- 2 A repository of climate and hidrology data
- 3 Conceptual Schema
- 4 Database Loading
- 5 Conclusions

# Outline

- 1 Motivation
- 2 A repository of climate and hidrology data
- 3 Conceptual Schema
- 4 Database Loading
- 5 Conclusions

# Outline

- 1 Motivation
- 2 A repository of climate and hidrology data
- 3 Conceptual Schema
- 4 Database Loading
- 5 Conclusions

# Outline

- 1 Motivation
- 2 A repository of climate and hidrology data
- 3 Conceptual Schema
- 4 Database Loading
- 5 Conclusions

# Outline

- 1 Motivation
- 2 A repository of climate and hidrology data
- 3 Conceptual Schema
- 4 Database Loading
- 5 Conclusions

# Motivation

- For decades countries have collected climate data. The proliferation of sensors have made these data grow tremendously.
- Meteorologists and researchers model climate data and try to predict events which pose risks to populations.
- The basis of these analysis are climate data.
- In each country several agencies collect data. Each agency uses different formats and different protocols for collection and dissemination of data.
- These diversity of climate data sources makes it difficult for users to access the data.

# Motivation

- For decades countries have collected climate data. The proliferation of sensors have made these data grow tremendously.
- Meteorologists and researchers model climate data and try to predict events which pose risks to populations.
- The basis of these analysis are climate data.
- In each country several agencies collect data. Each agency uses different formats and different protocols for collection and dissemination of data.
- These diversity of climate data sources makes it difficult for users to access the data.

# Motivation

- For decades countries have collected climate data. The proliferation of sensors have made these data grow tremendously.
- Meteorologists and researchers model climate data and try to predict events which pose risks to populations.
- The basis of these analysis are climate data.
- In each country several agencies collect data. Each agency uses different formats and different protocols for collection and dissemination of data.
- These diversity of climate data sources makes it difficult for users to access the data.

# Motivation

- For decades countries have collected climate data. The proliferation of sensors have made these data grow tremendously.
- Meteorologists and researchers model climate data and try to predict events which pose risks to populations.
- The basis of these analysis are climate data.
- In each country several agencies collect data. Each agency uses different formats and different protocols for collection and dissemination of data.
- These diversity of climate data sources makes it difficult for users to access the data.

# Motivation

- For decades countries have collected climate data. The proliferation of sensors have made these data grow tremendously.
- Meteorologists and researchers model climate data and try to predict events which pose risks to populations.
- The basis of these analysis are climate data.
- In each country several agencies collect data. Each agency uses different formats and different protocols for collection and dissemination of data.
- These diversity of climate data sources makes it difficult for users to access the data.

# The Project (started in 2006 with a grant from FONACIT)

## General Objective

To build a repository of climate and hidrology data for epidemiological and environmental risks management.

## Specific goals

- To recover historic data from analog media (paper bands, forms, among others).
- To store the data with quality annotations.

## Participants

- Seven laboratories at Universidad Simón Bolívar (USB) and Universidad Central de Venezuela (UCV).
- Multidisciplinary: statisticians, ecologists, computer scientists, geographers. (Aprox. 12 researchers)

# A repository of climate and hidrology data

## Our approach

- To design and implement *a database* to store climate and hidrology data.
- Instead of storing the collected data as disconnected time series, each kind of data will have a structure to hold it and their relationships to other data are made explicit.

# Methodology

- 1 Requirement Analysis
- 2 *Universe of discourse* specification
- 3 Conceptual schema development
- 4 Translation to a logical data model (relational)
- 5 Implementation using a DBMS (PostgreSQL)

# *Universe of Discourse*

## In general

Describes the relevant objects in the domain and their relationships.

## In particular, for the repository

- Covers a broad collection of concepts in the domain of meteorology and epidemiology.
- Its specification follows the recommendations of the World Meteorology Organization (WMO).
- Based on the experience of researchers with more than 25 years of experience in the collection and analysis of climate data.

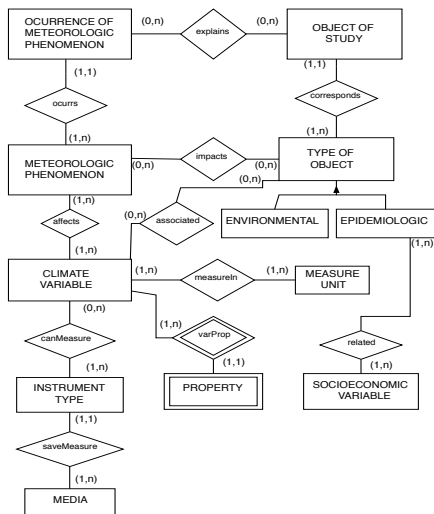
# Conceptual Schema for the repository

- Used the Extended Entity-Relationship Data Model (ERE)
- Developed in three stages and subschemas:
  - 1 Metadata Subschema
  - 2 Inventory Subschema
  - 3 Measurements Subschema

# Metadata Subschema

- Climate and hidrology variables and their properties
- Meteorological phenomenon
- Socioeconomic variables
- Objects of study, could be environmental (landslides) or epidemiologic (malaria or dengue cases).

# Metadata Submodel Diagram



# Inventory Subschema

- Data collecting stations
- Institutions
- Climate and socioeconomic variables for which there is data recorded in the repository
- Products developed by applying procedures to sets of data

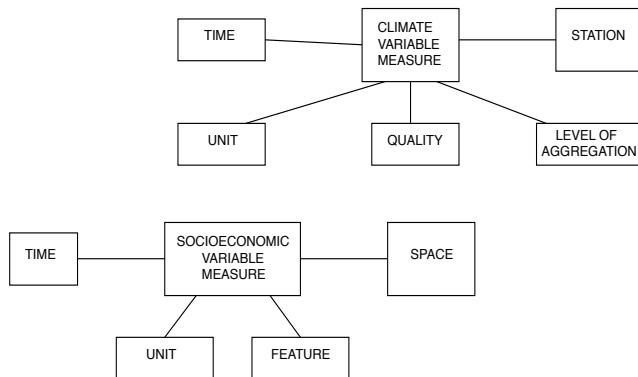


# Measurements Subschema

Climate variables measurements are modeled with a star schema.

- For climate variables like temperature, rainfall, humidity, there are many years of data from approximately 2000 stations nationwide.
- Each measurement was collected at a *station*, at some *time*, in some *units*. The measurement can be original as captured by the instrument or have been aggregated or processed jointly with additional data, this is reflected in the *level of aggregation*.
- A unique feature of this schema is data each measurement can be described by a quality annotation.

# Measurements Submodel Diagrams



# Integrated Schema

- The three subschemas described were integrated into a single global schema.
- A subset of the integrated schema was implemented in PostgreSQL.
- The big challenge was (and continues to be) the data loading.

# Characteristics of the data collected

- 1 The main agencies that collect data in Venezuela are: MARN (Minister for Environment and natural Resources), FAV (Air Force), Edelca (Electric Company on the Caroní River), INIA (National Institute of Agriculture).
- 2 Variety of media and formats on which data are collected:
  - data from paper bands (bandas);
  - special formats (planillas) where data are written by hand;
  - notebooks, another special form, hand written (climatological notebooks);
  - digital files (flat, excel or .dbf) summarized data from sensor stations with handwritten comments;
  - curves (handwritten as summaries of station data);
- 3 Varying temporal scales (hourly, daily, weekly, monthly).

# Database Loading

- Loading data into the metadata and inventory subschemas is relatively easy.
- But the station data has some inconsistencies and incompleteness. To validate these data properly, a physical inventory of stations needs to be performed.
- Focus on two main media: paper bands and digital files.
- The paper bands are collected from MARN offices all over the country. Many of these bands have been lost. We are trying to recover at least 30% of them.
- Digital scanning of the bands is manual, but the vectorization is a separate project. The repository will also contain the digital image of the band.

# ETL Procedures (the most challenging!!)

- For the data contained in digital files we built loading procedures.
- The files provided by each agency are analyzed to match the format with the database structures.
- An automatic procedure is constructed to parse the original file and generate the appropriate flat file or loading script.
- Each type of file from each provider institution requires a different procedure.
- The procedure is created once and (hopefully) will be used many times.

# The *Argus* Application

Another product of this project was the development of a Web 2.0 application, *Argus*.

- Front-end to the repository.
- Provides services over the data, for example a station georeferencing mechanism.
- Platform for transforming a product into a service.

The database and its companion *Argus* application constitute a digital library system.

It addresses the five cornerstones: content, user, functionality, quality and policy.

# Conclusions

- In this project we have designed a database to structure and store climate data.
- Instead of asking the data providers for a single, standard format, we accommodate data with diverse characteristics into a centralized repository.
- The database approach provides a mechanism for controlling the data that enters the repository and allows the addition of quality annotations.
- This platform is being evaluated by the governmental agency that is trying to centralize the climate data of the country.

# Thank you, gracias

Questions?

Soraya Abad-Mota  
abadmota@usb.ve