J2.4 MI^3: the NCDC's Master Station History System

Jeffrey D. Arnfield *
NOAA National Climatic Data Center, Asheville, North Carolina

1. INTRODUCTION

NOAA's National Climatic Data Center (NCDC) is the world's premier archive of weather and climate data. The data currently archived under the NCDC's stewardship span centuries, countries and continents, and are collected from numerous domestic and international agencies and observing systems. These data are vital to research, commerce, public health and law, and must be preserved and accessible for many years. This massive archive now exceeds two petabytes (two million gigabytes) and is expected to grow tenfold in four years.

To ensure the data's long term viability, an increasingly rich body of data about the data – metadata – must be also acquired, maintained and made accessible. Models and space-based systems play an increasing role in meteorology and climatology, but much of the data is observational, collected at discrete terrestrial observing sites or stations. Thus, observing station configuration over time is a critical component of these metadata.

As part of its Climate Database Modernization Project (CDMP) the NCDC has developed and implemented the MI3 (Metadata Integration and Improvement Initiative) Station Information Management System, a flexible, extensible system designed to meet the challenge of managing increasingly detailed station metadata from a growing number of networks.

Other papers (Arnfield, 2000; Arnfield, 2001) have been written describing the system. These include a more detailed description of the baseline state of station information management at NCDC, examples of the database normalization process, and a detailed discussion of some of the techniques used for date management and change journaling. The reader is invited to refer to these papers for additional information.

The MI3 system has now been operational for two years. Following a brief overview of the system's requirements and goals, this paper examines its organization, user interface and integration with other systems. It considers how well the system has achieved its goals so far, as well as challenges encountered and lessons learned during development and implementation. Lastly, the paper discusses some key challenges regarding the production, management and delivery of station metadata that must be addressed within the climate and meteorological community.

* *Corresponding author address:* Jeffrey D. Arnfield, National Climatic Data Center, Active Archive Branch, 151 Patton Ave, Asheville, NC 28801; e-mail: Jeff.Arnfield@noaa.gov.

2. BASIC REQUIREMENTS

NCDC needed a repository that could initially accommodate all stations contributing data to its archive and make the information available to all users, with potential to accommodate additional NOAA and non-NOAA stations.

The subject was considered in detail while developing requirements for the United States Climate Reference Network (USCRN). Further analyses of current systems, data sources and strategic direction, refined through subject matter expert workshops and surveys, yielded some general requirements:

- Accommodate all information contained in the existing Station History Information Production System (SHIPS) and produce the same production reports.
- Accommodate any fixed-position stations whose observations are archived at NCDC.
- Accommodate new networks and data programs involving similar stations, including those outside of NOAA, without modification of the database or user interface.
- Accommodate stations that participate in multiple networks.
- Accommodate highly detailed station information, including digital photographs, instrumentation, calibration records, observing schedules and data transmission protocols.
- Accommodate new instances of specific information types, such as phenomena observed, station identifiers or geographic regions, without modification of the database or user interface.
- Provide a means of defining and recording new types of information for a station.
- Provide a web-based user interface to permit location-independent access to station information.
- Provide flexible search tools to enable users to find stations by any of its names, various identifiers, geographic descriptors, parameters measured or network.
- Provide read-only access to all station details to all users.
- Permit restriction of such privacy-sensitive information as observer name and address on a user-by-user basis.
- Enable users to maintain only a specific group of stations.

- Permit easy correction of historical records, with minimal data duplication.
- Track the sources of information used to update the database.
- Develop an automated ingest process for station information where a suitable source is available.
- Permit additional, logically related metadata, such as data inventories, to be added or linked to the system at a later date.
- Permit other web-based systems to invoke the user interface for a specific station, providing access to station details with minimal effort and without data replication.
- Image, index and present existing paper forms to permit immediate access regardless of physical location.

The recent NOAA Integrated Surface Observing System (ISOS) project and the Global Earth Observing System of Systems (GEOSS) project have focused on the need to smoothly integrate observations and stations from diverse observing systems.

While MI3 system requirements and development predate these important projects, these and other initiatives reinforce the need for the flexibility designed into the MI3 system. MI3 will be evaluated in light of these initiatives to identify and address any shortcomings.

3. DATABASE DESIGN

MI3 uses a relational database design of about 130 tables, and is implemented use the Oracle relational database management system.

Each detail about a station may vary independently of most all others, and this fact complicates not merely date management but data management in general.

One approach to maintaining iterative versions of the station's configuration is to create a new record for the station in all tables each time a value in any table changes. This sequential snapshot approach can simplify queries, and is often a useful technique for submitting new station information. However, it involves a great deal of data redundancy and greatly complicates historical correction, since changing a single longitude may involve updating dozens of records.

Instead, the MI3 database uses a highly normalized design. Each record contains a beginning and an ending date for the information it contains. A new record is created in the table only if a value changes for a column in that table. Screens and reports are then populated by joining records from multiple tables rather than querying monolithic records

A result of these varying dates of validity for each of a station's records in multiple tables is that the beginning and ending dates in a time-series view of the station's configuration may change whenever a new field is added to the view. Additional complications occur when a table may not contain data for a given period. The context date pair technique described by Arnfield and Shears in 2000 was developed to provide a means of managing these dates, and has proven effective in developing the user interface, custom data views and reports.

In a relational database, a record in one table may have a parent/child relationship a record in another table, established using a common key field. Referential integrity constraints ensure that only one unique parent record is referred to in such a key relationship, and that the parent record in a table cannot be deleted if another record refers to it.

While maintaining separate dates in each record for a station in each table has minimized data duplication, it has created potential issues in maintaining data integrity. Logically, the child record must fall within the date range of the parent without exceeding it; however, this sort of range validation cannot be defined as part of the key relationship that referential integrity relies on.

Several peer level records may exist, sharing the same parents. It may also be necessary to ensure that their begin/end dates do not overlap. Again, referential integrity checks do not ensure logical consistency in this situation.

Both situations were addressed in the user interface, but it is still possible to create date problems by directly updating the database. If database-level updates were performed outside the control of the user interface and accompanying stored procedures, we'd need to protect against such parent/child and peer/peer date conflicts. One approach would be to use database triggers invoked each time a record is update to validate the date ranges and prevent the update if there are conflicts. Another approach would be to develop automated audit queries to check for such conditions so that they could be corrected. While triggers would provide more certain protection, performance considerations might make the audit and correct technique preferable.

Many systems and processes rely upon metadata. While we must know when something changed at the station, we also must record when we learned about the change and recorded it. It may be that data processed using out-of-date values should be reprocessed, or at least interpreted differently. MI3 tracks the dates that information is changed and journals previous values; it also tracks sources of information used in updating records. In the future, a view and report showing individual value changes to each station by time period may need to be developed.

4. USER INTERFACE

The MI3 system provides both a simple and an advanced query interface. The simple query permits the

user to search for stations by name or identifier. Optionally, the user may restrict by station type.

A more advanced interface permits the user to search by specific phenomena observed, by date range, various geographic descriptors, within a latitude and longitude range, and various other conditions.

Station names change over time, and at any given time a stations may be known by multiple names. For that reason, MI3 searches all names and aliases recorded for a station. It also supports searching names using "starts with," "contains," "ends with," "exactly equals" and "sounds like" options, greatly increasing the user's chances of finding stations of interest.

A basic grid of search results is presented, giving the user a quick overview and an opportunity to drill down for additional detail.

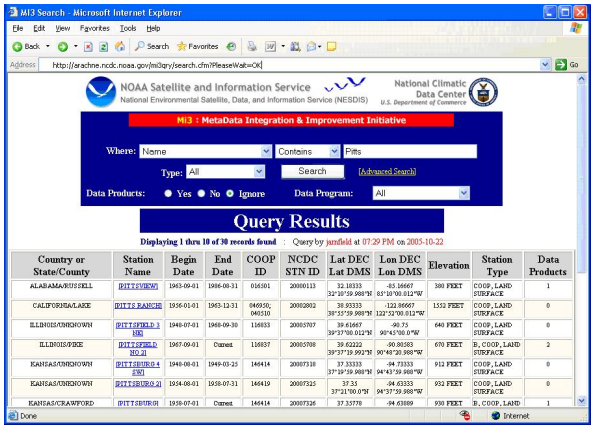

Figure 1: Query results window

MI3 station information is organized into broad subject areas to help users find information as well as easily partition the system for development. The general subject areas originally identified for the USCRN metadata were used as a starting point:

- identity
- updates
- location
- involved parties
- data programs
- data products
- equipment
- phenomena
- map
- remarks
- utilities and administration

Within the system a consistent header appears at the top of the page, with a series of folder-style tabs based on the subject areas affording the user easy navigation.

A time-series summary of each subject area's data is presented in a grid. The user can click on an individual grid row to see a more detailed form view. Users with maintenance privileges can also access the maintenance screens from either the grid or form view, providing a fairly seamless transition from research to correction.

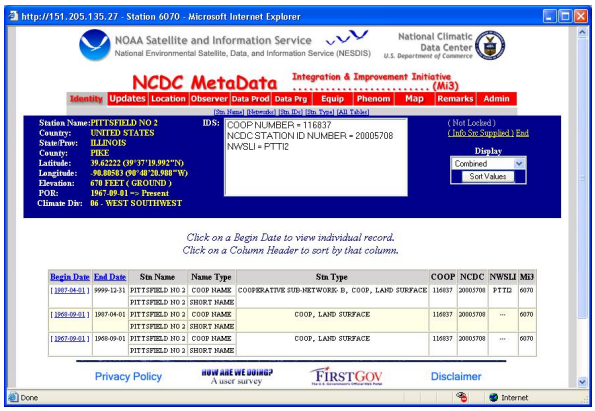

Figure 2: Station details window, identity grid

Balancing flexibility, ease of use and development constraints is often a challenge. While the context date pairs were quite successful in reports and views, they do not resolve the intricacies of varying effective dates within the maintenance interface. A variety of constraints forced the first iteration of the user interface to provide only table-level maintenance rather than integrated screens for each subject area. This was judged acceptable for the short term because the greatest volume of station information would be automatically ingested from the National Weather Service's Cooperative Station Service Accountability (CSSA) system.

While serviceable and permitting the system to be implemented, this approach was cumbersome and required too much user training. A second implementation of the system includes integrated maintenance screens for each subject area, and has been well-received by users.

5. SYSTEM CONTENT

MI3 was initially populated with legacy station information from the existing SHIPS station history system, yielding roughly 30,000 stations. The most aggressively maintained details had been for the roughly 27,000 stations in the Cooperative network. It is not unusual for a Cooperative station to have existed for many decades, and to have undergone more than a dozen significant configuration changes over time, resulting in a large number of detail records.

All commissioned ASOS stations were also part of the initial load, although the records contained less detail. About 5000 additional stations, mostly historical, were from various networks, contained sparse detail and had not been well maintained, due in part to a lack of available information.

During conversion of legacy station information many discrepancies were found, mostly involving the earliest and latest date that each fact about the station was valid. Thousands of records were researched and corrections were made. This time consuming process greatly improved the accuracy of the data.

Since becoming operational, all 886 commissioned ASOS stations have been added to MI3. NCDC has also added more than 530 AWOS sites, all 160 NEXRAD stations, all 75 commissioned USCRN stations and hundreds of additional Coop stations to the system.

While the SHIPS system tracked a station's participation in two particular data products, MI3 contains participation information for 12 datasets. Currently only dates of participation are noted, but additional detail will be added in the form of data inventories. This enhancement is discussed later.

6. ONGOING INFORMATION INGEST

There is no national content or format standard for exchanging station metadata, and not all networks provide this metadata in machine-readable form. Thus, either information must be entered manually, or automated ingest programs must be developed on a data source by data source basis.

Because the NWS Cooperative network provides the largest volume of station information annually and is now entered and quality controlled using an Oracle based system called CSSA, we developed an automated ingest routine to process Cooperative station history submissions.

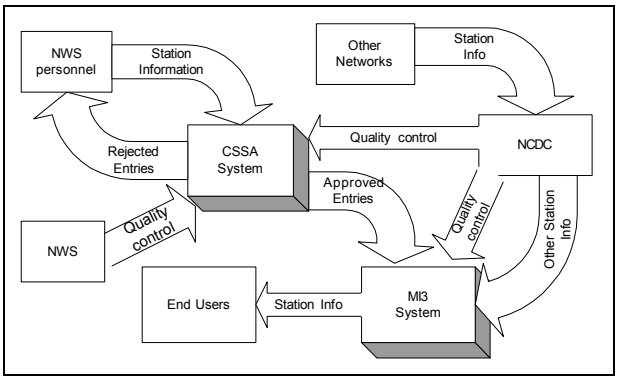

Figure 3: Station Information Flow

Information from other networks is currently entered manually, although an automated ingest process is being developed for stations appearing in the World Meteorological Organization's (WMO) Publication 9, Volume A. As additional station information sources are identified that are suitable candidates for automated ingest, based on importance to our mission, data volume, update frequency and consistency, we will develop other automated ingests.

Lack of accessible, comprehensive, timely station information complicates station metadata maintenance. Some operational networks require only current station information, and historical values may not be preserved. They may use several discrete systems for different aspects of station management, leaving no single, complete source from which to extract station metadata.

Standards and best practices for station metadata content and format are slowly evolving, but current Federal Geospatial Data Committee (FGDC) and International Organization for Standardization (ISO) standards to not adequately facilitate exchange of detailed station metadata. As common standards emerge and are adopted, MI3 will be made to exchange station metadata using a common format.

A separate database of more than 18,000 historical Cooperative stations will be integrated into MI3 as part of a separate project. More than 5000 of these stations do not currently have entries in MI3, and this load will significantly extend our content. While loading the data is fairly straightforward, the database often contains entries with dates that overlap and values that conflict with MI3. A manual review and resolution process will be necessary to complete the integration.

Much detailed historical information for stations exists in document rather than digital form. While many of these documents have now been imaged for easy access (discussed in a later section), the information still needs to be entered in to MI3. This entry will occur over a period of years.

7. CURRENT USERS

Most users access MI3 via the guest account, which permits viewing all details except observer information. Users are NCDC, the National Weather Service, Regional Climate Centers (RCCs), State Climatologists, researchers at universities and a variety of private industries. Key reports from MI3 drive Cooperative data ingest and publication at NCDC, and are used in inventory production. Several NCDC systems and projects, including Climate Data Online (CDO), NEXRAD Inventory Visualization and Health of the Network, either link to MI3 or use information extracted from it. MI3 extracts are also a basic source of station information for the Applied Climate Information System (ACIS) developed with the RCCs.

8. DIRECT ACCESS BY OTHER SYSTEMS

Web-based systems can easily instantiate an MI3 window of station details for a specific station by adding a Common Gateway Interface (CGI) GET-style parameter list to the web uniform resource locator (URL) used to access the MI3 system. A station identifier is specified, along with a code for the identifier type.

There is no universal identifier for all stations from all networks, and most data programs and datasets use a single type of identifier, such as a Coop ID or a call sign, to identify their stations. MI3 permits searching by any of a station's identifiers.

Sometimes identifiers have been assigned to more than one station over time. If more than one station is found with the specified ID, the query results list is presented for those stations and the user can select from that list.

The basic syntax is straightforward to developers familiar with web protocols:
http://arachne.ncdc.noaa.gov/mi3qry/displaystation.cfm?idtypeabbr=ICAO&idvalue=KAVL

The value for "idtypeabbr" may be one of the following:

- "ICAO" (International Civil Aeronautics Organization call sign: 4 alphanumeric characters)
- "WBAN" (Weather Bureau Army Navy ID: 5 digits)
- "COOP" (Cooperative station ID: 6 digits)
- "FAA" (Federal Aviation Administration Call Sign: 3 characters)
- "WMO" (World Meteorological Organization Index Number: 5 digits)
- "NWSLI" (National Weather Service Location Identifier: 3-5 alphanumeric characters; inconsistent for older stations)
- "GOES" (Geosynchronous Orbiting Satellite: format varies; used for GOES transmission, and currently entered only for CRN stations)
- "CRN" (The internal Station Index identifier used for CRN network station management)
- "NCDCSTNID" (NCDC Station ID: 8 digits; assigned by NCDC for internal station management purposes)

Direct access to the Oracle database is possible using Oracle's SqlNet, but for reasons of security, support and performance such access is currently limited to internal NCDC users. Due to the number of tables and interrelationships, queries may be quite complex. In the future, specific views of the data will be developed to simplify ad hoc queries, help ensure valid results and minimize the performance impact of such queries.

9. UNDER DEVELOPMENT

The MI3 system is being actively enhanced. Several key functions are currently being developed that will increase both the functionality and content of MI3.

Support for management and display of digital images will be added to MI3 in early 2006. Several years ago digital photographs were taken of roughly 211 Automated Surface Observing System (ASOS) sites as part of a project to document wind exposure in support of tropical cyclone research. Some networks, such as the USCRN, now make extensive use of digital photography to document station siting and exposure. NWS is now using digital photography in some regions to document Cooperative stations, and its use is expected to spread. Digital maps and satellite imagery must also be accommodated. Enhancing MI3 to accommodate such information will provide station managers and researchers with new opportunities to thoroughly document a station's configuration and environment.

Once a user locates a station of interest, the next question is often "what data are available from this station?" NCDC uses a system called PC CliServ to present detailed data inventory information, but the system is accessible only within NCDC. An initial load from CliServ permitted us to note dates of participation in a dozen datasets in MI3 without structural or programmatic modification.

Development is underway to incorporate more detailed inventory information to MI3, using a flexible database structure that will accommodate multiple levels of inventory, both at the station and the geographic region level. The database and interface modifications should be completed in early 2006.

A separate Integrated Inventory Development project is now underway at NCDC to ensure that detailed and timely inventories are produced and available for all datasets archived at NCDC. When completed, it should significantly improve the quality and detail of NCDC's data inventories.

To permit other users to enter new groups of stations without being able to modify existing stations, a new layer of security is being developed. This will permit users outside of NCDC to enter as well as query station information.

10. DEVELOPMENT PLANS

In the coming year, MI3's reporting and data export capabilities will be expanded. A comprehensive station-level snapshot report will be developed, as well as a flexible report and export engine with a variety of output options.

GIS access to data is almost de rigueur for modern systems. Rather than develop a separate GIS interface to MI3, we will use existing NCDC GIS systems to help users locate and access stations, with MI3 providing detailed information for each station.

Entering station information is labor intensive, and manual processes always have potential for error. A quality control workflow system will be developed to accommodate USCRN station metadata, along with an automated ingest of the approved information into MI3. Such a system will permit more timely and accurate update of station information. The basic approach will

be extended to accommodate other networks as interest warrants and resources permit.

11. ANCILLARY ACTIVITIES

Much of NCDC's detailed station information was available only in paper form. A secondary project was conducted to review, collate and image paper station history forms and make them available via a web-based document management system. This has been completely successful, and having the forms readily accessible via the web has greatly increased productivity during research of anomalies. Privacy concerns surrounding the volunteer observer address and phone number information on the forms limit general public access to the information.

Since lack of physical access to data forms is no longer an issue and the MI3 system can be accessed anywhere via the web, a future effort to manually enter all data from these forms into the MI3 system can involve a partnership with contractors, regional climate centers or state climatologists.

NCDC has also developed a series of web pages, available at http://mi3.ncdc.noaa.gov, to provide information about MI3 and direct access to the system and its standard report products. Documentation for the system is also available there.

12. CONCLUSION

While development is still underway, the MI3 system has been successful in making a richer body of station information available to the entire user community, and has been able to accommodate new station networks without structural modification to user interface or database. Enhancements currently under development will permit a broader variety of metadata, including digital imagery, to be captured for a station, and will further its progress toward the long term vision of integrating station metadata more closely with collection level and inventory metadata.

13. REFERENCES

Arnfield, J. D., 2000: A Flexible System To Manage And Query NOAA Station History Information, American Meteorological Society, 17th Conference on Interactive Information Processing Systems, Albuquerque, NM, Jan 14-18, 2001; p468-47

Arnfield, J. D., Gary Shears, 2001: Station History Database Architectural Techniques, American Meteorological Society, 18th Conference on Interactive Information Processing Systems, Orlando, FL, Jan 12-18, 2002

Arnfield, J. D., et al, 2000: U. S. Climate Reference Network, Part 4: Metadata. American Meteorological Society, 12th Conference on Applied Climatology, Asheville, NC, May 8 – 11 2000

Kevin Robbins, et al, 1999: Metadata for the Unified Climate Access Network. Third IEEE Computer Society Metadata Conference, Bethesda, MD April 6-7 1999

Vironst-Lazar, A. K., Kevin Robbins, 1999: Advancements in the Integrated Management of Site Metadata for Multi-Agency Weather/Climate Data Networks. Third IEEE Computer Society Metadata Conference, Bethesda, MD April 6-7 1999