9. MST RADAR DATA-BASE MANAGEMENT
(Keynote Paper)

V. B. Wickwar

SRI International
Menlo Park, CA 94025

The management of data is an important topic and one that has had considerable attention during the last few years (e.g., CODMAC, 1982; LIDE, 1981; JENNE, 1981; ROEDERER, 1981; USRA, 1978). Indeed, for MST radars, there are several reasons for considering it. The first is a very practical consideration, the large amount of data that are acquired by an MST radar, whether operating continuously or in campaign mode. Another is the desire to exchange data among scientists at the several radars located around the world. Now that the MST technique is maturing, there is also a growing desire by researchers using other techniques to acquire correlative radar data and a desire by theoreticians to compare theory and observation. Thus I believe that this is an appropriate time to discuss the management of MST data.

However, I do not want to step forward as an expert in the area of data management. I was, nevertheless, greatly involved as a catalyst in setting up an incoherent-scatter radar data base at NCAR. Let me emphasize the word "catalyst" because setting up such a data base has to be a collective effort of scientists involved in obtaining and using the data in question.

What I would like to do in this article is review some ot the questions and concerns that were involved in setting up that data base. That experience should be relevant to the present concerns because of the similarity of the techniques and the overlap in the scientific community interested in data from the two techniques.

The first question is the purpose of the data base. Is it to bring together all the data from one type of instrument, or is to bring together all the data from any source that pertains to a given scientific problem? The solution we chose was a blend of both, but it tended to emphasize the former. We felt that certain portions of the radar data could make particularly valuable contributions to the scientific field, and therefore we recommended including them. But for the data to be useful, we felt certain geophysical parameters should also be included. While we did not recommend the inclusion of other major data sets, we did recommend that that option be left open and, similarly, that the option to obtain data from other data bases not be precluded.

The rationale was that keeping the data base limited addid to the likelihood of having a responsive facility and one that would greatly complement the radars over the years. We even hoped that the facility would become a center of excellence for using the radar data. In another direction, we believed that this type of data base could contribute easily and greatly to short-term CDAWS-like exercises where many varieties of data are brought together to study a particular event or problem.

The next question we considered was the nature of the data base. Should it be a distributed system or a centralized one? In this age of networking and decentralization, we chose a centralized facility. This came about because of two realities: (1) the present cost of having a network linking radars in different countries or, even worse, on different continents, and (2) the expense of having all the data on line all of the time or even portions of the data shifted around on a fixed schedule. It also came about for positive reasons. We felt there would be more uniform quality control if all the data passed through a common gateway. We felt that a central facility would be in a better position to train and help data-base users, to be able to assemble catalogs of observations, etc.

Another aspect of the nature of the data base is whether it should be on line or not. While having the data on line would be very nice, it is expensive; and it was not clear that it was necessary. The data could be kept on magnetic tape and then promoted to disks when requested. It could then be left on disk for a few days in case it was looked at again. The delay in promoting a tape would not affect most users. However, it was felt that something small and very useful for selecting data, such as a catalog, should be kept on line for easy browsing.

Yet another aspect of the nature of the data base was the nature of the data-base management system. Almost all computer systems have data-base management systems. However, while they may be excellent for keeping track of mailing lists, inventory control, or customer accounts in financial institutions, they are not well adapted to scientific data. Among other reasons, the data formatting may vary very greatly from radar to radar, or even for a single radar from one experiment to the next. Another is that relationships may exist with a series of points instead of on a one-to-one basis. However, the data can be handled simply with a far simpler file management system.

Having examined the nature of the data base, the next question was what data to include. This had two major aspects. The first was the level of the data. For example, there are the raw data that are recorded on tape for subsequent reduction. They are very radar-specific and comprise the bulk of the data. However, these data can only be reduced to geophysical parameters at the radar site. Thus, we felt that they should be kept at the radar location and not, at least initially, made part of the data base. However, magnetic tapes do deteriorate with time, and so it seems advisable that those tapes should be copied on to new tapes or into new storage media in the future if they are to be saved. The next level of data consists of geophysical parameters derived directly from the power and spectral shape as well as from combining line-of-sight velocities into vectors. Because at this level the parameters are independent of the radars and are used directly in scientific studies or used to derive a higher level of parameter, they are the ones submitted to the data base. Finally, the next higher level of data could then be derived in a uniform manner from data from any of the radars. They would be best derived by the users with their own programs or those available at the central facility. So, essentially, our rationale in deciding how to treat the various levels of data was based on what could best be done at the most appropriate location.

The second aspect of what data to include had to do with whether the data base should include data from all experiments or from only a subset. The decision was to include all long experiments that observed standard ionospheric parameters. The rationale was that these longer experiments would be the most usable by those other than the original observers. Similarly, it was felt not appropriate to include studies of special parameters or of new altitude regions. Often such studies are experimental in nature, i.e., aimed at extending the technique, or are of limited interest to the wider scientific community.

Having decided on what data to include, our next area of concern was with what tape format to use for transferring data. The aim was to adopt a format that was well specified yet versatile, one that would handle any type of radar data -- present or future -- plus other types of correlative data. To do that the format also had to be largely self-documented. The result was a format with three types of records. The first is a protocol record that uses ASCII characters to describe the experiment and data. The second is a header record that identifies the radar, date, time, pointing directions, etc. Following the header is the third type, a data matrix of values and identifiers. The second and third records are written in twos-complement 16-bit integers. While the machinery was moving forward to establish the data base, the three radars involved in Project MITHRAS (Chatanika, Millstone Hill, and EISCAT) implemented the tape format and used it for exchanging data.

Although the foregoing considerations have largely been concerned with technical aspects of the data base, there are some very important human considerations that must not be forgotten. For the data base to be useful, the users must have confidence in the quality of the data or have a mechanism for assuring themselves of that quality. Similarly, if they do not understand how some of the geophysical parameters were derived or what their limitations are, they need a mechanism for finding that out. Analogously, the scientists working with the radars need a compelling reason for reformatting their data and submitting it in a timely fashion to the data base. Although data acquired with public funds usually have to be made available, it was not felt that coercion was the best method to ensure data quality. Instead, in recognition of the effort put in by the scientists working with the radars, it was felt that the "carrot" was the best approach. A set of "Rules of the Road", such as those pioneered by the Atmospheric Explorer team, was developed to set out the responsibilities of the providers and the users. The critical point was that, in exchange for providing the data in a timely fashion and for answering questions about quality and derivation, the appropriate scientist working with the radar could be offered coauthorship on resulting papers and reports.

While it is implied in the foregoing paragraphs that there would be appropriate people at the facility housing the data base who would be entering the data, the committee felt that for the data base to prosper and evolve, there had to be at least one scientist associated with it who would, among other things, use the data. He or she would therefore be intimately aware of the programs to access the data, plot, derive higher level parameters, and compare them to advanced models. He or she would also participate in the development of those types of software. The efforts of this person would be supplemented by outside scientists and visitors contributing additional software.

The deliberations that led to the establishment of the data base were conducted by a committee carefully selected to include representatives of each radar, the user community, and experts in data-base systems. It was strongly felt that some such committee had to be established on a permanent basis to assess how the data base was developing and to guide its evolution.

In conclusion, the management of data is exceedingly important, particularly in those disciplines that have to rely on extensive observations instead of controlled experiments. The solution of putting data in a "data base" is a very appealing one. However, I hope my descriptions of some of the issues that came up during the meeting on the formation of an incoherent-scatter data base have convinced you that the doctor cannot write a simple prescription for "one data base". Considerable thought, discussion, and active involvement are required of the whole community. Let us continue to think, discuss, and become involved during the remainder of this workshop and beyond.

ACKNOWLEDGMENTS

The other participants in the Incoherent-Scatter Radar Data Base Workshop were D. Alcayde, J. Baron, R. A. Behnke, O. de la Beaujardiere, B. A. Emery, C. Gonzales, T. Hagfors, J. M. Holt, R. L. Jenne, D. H. Joseph, M. C. Kelly, L. A. Lee, W. M. MacIntyre, E. S. Oran, A. D. Richmond, R. G. Roble, R. W. Schunk, P. Shames, J. Silen, W. Swartz, and R. F. Woodman. Support for participation in this workshop was provided by SRI International.

REFERENCES

CODMAC (Committee on Management and Computation) (1982), Data Management and Computation, Volume 1: Issues and Recommendations, Space Science Board, National Academy Press, Washington.

Jenne, R. L. (1981), Strategies to develop and access large sets of scientific data, In report of conference on "Frontiers in Data Storage Retrieval and Display", National Geophysics and Solar-Terrestrial Data Center, EDIS/NOAA, Boulder.

Lide, D. R., Jr. (1981), Critical data for critical needs, Science, 212, 1343-1349.

Roederer, J. G. (1981), Considerations in the development of a national geophysical data policy, EOS, 62, 569-570.

USRA (Universities Space Research Association) (1978), Proposal for a national data center to promote data analysis and theoretical research in solar terrestrial physics, Final Report, USRA Data Analysis Panel, Houston.