Chopo Ma: Data Longevity Beyond 2010, IVS 2004 General Meeting Proceedings, p.97–99
http://ivscc.gsfc.nasa.gov/publications/gm2004/ma2

Data Longevity Beyond 2010

Chopo Ma

NASA Goddard Space Flight Center
e-mail: cma@virgo.gsfc.nasa.gov

Abstract

The current cache of S/X-band geodetic/astrometric VLBI data accumulated since 1979 is ∼5 million observations and is increasing by ∼300 000 observations per year. The long time interval and access to all such VLBI data for re-analysis have contributed to their usefulness for the terrestrial and celestial reference frames, Earth orientation parameters, tidal and nontidal loading, and troposphere. While data access and integrity have been maintained through the Mark III data base system as storage devices and media have evolved, past transitions have been major projects. A new format and retention concept to ensure eternal archiving and access should make use of self-documentation, generalized media, network connectivity and multiple redundancy. Similarly permanent organizations or sequences of organizations are also necessary.

1. Introduction

The size of the current dual frequency geodetic/astrometric data set is ∼5 million delay and delay rate observables accumulated by various national and international observing programs since 1979. The data set is growing by ∼300,000 observations a year, and the rate is likely to rise as technical developments such as the Mark 5 recorder and e-VLBI improve the efficiency of data acquisition and correlation. The usual forms for data distribution and archiving of geodetic/astrometric data are Mark III data base files and NGS card images, both developed in the 1970s. These data files are available at various IVS data centers in North America, Europe and Asia. In addition to VLBI data used directly for geodetic and astrometric analysis, most correlators retain their raw correlator output and some also archive the fringe results. Since one of the strengths of VLBI as a space geodetic and astrometric technique is the ability to re-analyze the entire data set with improved modeling and estimation, it is essential to maintain data access and integrity indefinitely. An analogy is perhaps astronomical plates, which have had utility far beyond the lifetimes and intentions of the original investigators. This paper will concentrate on the delay/delay rate observables with some mention of the much more voluminous and varied correlator and fringe output.

2. Content

Compared to the number of bits recorded at the VLBI stations, the data used directly for analysis is quite limited, no more than a few Gbytes in the most compact binary form. The fundamental information is the output of the fringe finding process: various delays, delay rate and fringe phase along with their uncertainties as well as baseline, source, time tag and observing frequencies. Other outputs include phase calibration from individual frequency channels and a quality flag. Some information comes from individual stations such as cable length calibration and weather data. To reduce duplicate effort some primitive analysis results have been retained in the

Mark III data base files such as ambiguity resolution, ionosphere calibration and editing. These entries are not strictly necessary and can be redone if necessary. Because of the data flow design of the Mark III analysis system, there is considerable other information in the data base file by the stage at which the primitive analysis results are included.

3. Format

The current Mark III and NGS formats have demonstrated their utility by their continued use but both are far from ideal. In particular the former holds much extraneous and redundant information while the latter lacks certain data and any flexibility. A new format is required that can make use of current computer speed, memory size and disk capacity while retaining and extending the desirable features of the current formats. The new format should be self-documenting in terms of data history, content and organization. It should be self-contained, not requiring information in other locations for data extraction. The self-documentation and self-containment could include embedding the relevant data extraction software within the data organization. The format must be expandable since new data types and dimensions are sure to arise. Although not as compact as binary, the fundamental format should be ASCII or another text type to allow examination and access at a primitive level if necessary.

4. Media

The media must provide two functions, on one hand active media for easy, efficient routine access and on the other passive media for physical security and integrity. For the foreseeable future the ready access medium is likely to be magnetic disks attached to computers or networks. This medium permits simple extension of the data set as more VLBI data become available and has sufficient capacity at low cost to keep the complete data set on line. Periodically the data set must be copied to media that can be physically removed and stored. Since the complete VLBI data set is not large and is unlikely to grow more rapidly than media capacity, the passive media copies will not be physically large or expensive. However, the history of information technology has demonstrated that such media evolve continually, sometimes leaving data stranded and unreadable. Although optical disks and solid state devices like flash memory may be the passive media for the intermediate future, hard copy, perhaps microscopic, might prove to be the most durable and secure. More important than the actual passive media are the regular archiving and cataloging of the complete data set.

5. Access

Easy, efficient access to the VLBI data set for both extension and retrieval is an essential element to ensure continuing use and improvement of the analysis results. On an everyday basis the data should be available via network from multiple, mirrored data repositories to assure fast, reliable retrieval. In addition to catalogs of the data holdings, metadatabases with such details as networks and sources should be maintained at the data repositories and at other relevant organizations. The VLBI data should also be included in larger data aggregations and any integrated space geodetic data system such as INDIGO (IERS, 2003).

6. Archiving

The archiving of the VLBI data set may present some particular difficulties because of the extended time frame during which the data will continue to be useful. At any given time, the existing data repositories must be responsible for backing up their active media, if only to prepare for the inevitable system crash. Provided the proper organization exits, the other data repositories can also assist. However, in parallel to this routine precaution complete archive copies on suitable passive media should be made on a regular schedule and distributed to other interested organizations such as the space geodesy services, geodetic agencies, astronomical observatories or the VLBI stations. Storage and cataloging of such archive copies can be very simple at these locations. The essential point is that these organizations should have a long term interest in the continued use of VLBI data even as some VLBI entities may cease their activities. By wide and continual distribution of the complete current data set, its longevity can outlast any individual component.

7. Correlator Data Management and Archiving

Several of the correlators that have processed geodetic/astrometric observations have retained both the correlator and fringe output. Because of the amount of data and absence of regular reprocessing, these data are not kept on active media and exist only at the original correlation facility. The data have migrated, sometimes incompletely, through various passive media as such media have evolved. Since the potential exists for refringing of the correlator output with better algorithms to decrease instrumental and ionosphere noise, such archives should be kept. However, the management of these correlator archives will probably remain with the correlators although catalogs of these data should be maintained along with the catalog of the analysis data.

References

[1] IERS, IERS Annual report 2002, p. 14, 2003.