ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201909.00186

Postprint: A Radio Measurement Set File Generation Method Based on python-casacore

Authors: Sun Haomin, Deng Hui, Mei Ying, Wang Feng

Date: 2019-09-18T09:47:00+00:00

Abstract

MeasurementSet (MS) files have become a crucial storage file format in radio astronomy, gradually establishing themselves as the standard format for radio astronomy data storage, analysis, and sharing. They are supported by an increasing number of astronomical data processing software packages such as CASA and WSCLEAN, and are utilized in many radio telescope systems including ALMA and LOFAR. However, the MS format has seen limited domestic application for a long time, with technical specification documents being extremely scarce. This paper addresses the requirements of the SKA Engineering Bridging Phase by introducing the basic concepts, directory structure, and field design of the MS format. Building upon this foundation, we discuss the methodology for generating MS files using python-casacore to interface with the underlying casacore library, and encapsulate this functionality into the current SKA Algorithm Reference Library (ARL). The paper presents an example of generating MS files through simulated observations using ARL, and performs imaging on the generated MS files using CASA software. Through comparative analysis of the results, the correctness of the MS files is verified. The work presented herein provides critical support for subsequent SKA imaging experiments, observation simulation, and file storage, while simultaneously offering a valuable reference for radio astronomy data processing efforts both domestically and internationally, thereby fulfilling the requirements of the SKA Engineering Bridging Phase.

Full Text

MeasurementSet File Generation Method Based on Python-casacore

Haomin Sun[1], Hui Deng[12]*, Ying Mei[1], Feng Wang[12]

[1]School of Physics and Electronic Engineering/Center for Astrophysics, Guangzhou University, Guangzhou, Guangdong 510006, China
[2]Kunming University of Science and Technology, Kunming, Yunnan 650051, China
(*Corresponding author: denghui@gzhu.edu.cn)

Abstract

Measurement Set (MS) files have become an important storage format in radio astronomy and are gradually emerging as the standard for radio astronomical data storage, analysis, and sharing. This format is now supported by an increasing number of astronomical data processing software packages such as CASA and WSCLEAN, and has been adopted by numerous radio telescope systems including ALMA and LOFAR. However, the MS format has seen limited application in China, with very scarce technical documentation available. In this paper, we introduce the basic concepts, directory structure, and field design of the MS format in the context of the SKA engineering bridging phase. We discuss in detail the method for generating MS files by using python-casacore to call the underlying casacore library, and encapsulate this functionality into the current SKA Algorithm Reference Library (ARL). The paper provides an example of generating MS files through simulated observations using ARL, and verifies the correctness of the generated MS files through imaging with CASA software and result comparison. This work provides critical support for subsequent SKA imaging experiments, observation simulations, and file storage, meeting the needs of the SKA engineering bridging phase while also offering a valuable reference for radio astronomical data processing both domestically and internationally.

Keywords: Measurement Set (MS) files; Python-casacore; CASA

Introduction

Radio astronomy in China has experienced remarkable development over the past decade. The 21CMA telescope array has driven domestic low-frequency radio astronomy research, while the FAST telescope—once a candidate for the Square Kilometre Array (SKA) and later independently developed into the world's largest single-aperture telescope—has made significant contributions. Other domestic radio telescope arrays include the Tianlai array in Xinjiang, the Mingantu Solar Spectral Radioheliograph (MUSER) in Inner Mongolia, and the Chinese VLBI Network. These radio interferometric arrays have made important contributions to astronomical research, data processing, experience accumulation, and talent cultivation, laying a substantial foundation for China's participation in SKA.

Data storage is a fundamental requirement for radio astronomical observations. For a long time, FITS files have been the standard format for astronomical data storage, with UVFITS and FITSIDI formats developed specifically for radio

data storage. In recent years, Measurement Set (MS) files have gained increasing popularity and become an important storage format in radio astronomy, gradually emerging as the standard format for radio astronomical data analysis. MS is widely supported by astronomical data processing software such as CASA (the Common Astronomy Software Applications package) and WSCLEAN. However, MS format applications remain relatively rare in China, with limited documentation available in both Chinese and English. Consequently, domestic radio telescopes typically define their own raw data storage formats based on the characteristics of their receivers. For example, MUSER uses a bare binary format to save observation files, significantly reducing storage space. When data exchange is required, conversion software is used to transform the data into UVFITS or FITSIDI formats.

With the launch of the SKA bridging phase, the authors encountered the challenge of generating MS files during their involvement in testing and simulation work to interface with other mainstream astronomical data processing tools. Therefore, this paper systematically discusses the implementation of MS file generation using Python and python-casacore based on practical development needs, providing a valuable reference for subsequent work and data storage and sharing for other radio telescopes.

I. Concept and Definition of MeasurementSet Files

1.1 Basic Concepts of MeasurementSet Files

MeasurementSet files are a file format that follows the Radio Interferometer Measurement Equation (RIME) and can store raw radio astronomical data prior to calibration. Following the publication of the MS design standard, multiple astronomical software development teams, including the CASA team and the European VLBI Network, implemented the format in their code. Since CASA adopts the measurement equation as its fundamental calibration scheme, MS files naturally became the storage standard for radio observation data in CASA software. As CASA became the designated data processing and analysis software package for ALMA and VLA, MS accordingly became the default data format for ALMA and VLA data analysis. The raw data storage formats for ALMA and VLA are ASDM and SDM, respectively, and relevant software has been developed to convert data from these formats to MS.

An MS file is essentially a relational database whose format covers all conceivable use cases in radio astronomy, ranging from single-dish telescopes to simple interferometers composed of a few antennas, and up to large-scale interferometers with hundreds or thousands of antennas. MS adopts relational database modeling methods to reduce data redundancy by constructing primary and foreign keys. Data that appears multiple times (such as antenna data) is placed in separate database tables (subtables), which are then referenced in the main body of the database (the main table) through corresponding indexes (primary keys). In cases with two levels of subtables, the first level is referenced by

the main table, and the second level is referenced by the first level, enabling subtables to reference other subtables.

During MS file generation, most data—including interferometric visibility functions and/or single-dish total power measurements along with their timestamps—are stored in the main table, while most metadata resides in secondary subtables. The main table must generally include either a DATA column (for interferometric data) or a FLOAT_{DATA} column (for pure single-dish data), with one of these two columns required depending on the specific radio telescope.

1.2 MS v2.0 File Structure

The CASA software adopts MeasurementSet version 2.0 as its data format standard. In fact, MeasurementSet was formally defined in AIPS++ Note 191. To ensure data compatibility with CASA software, this paper also adopts MeasurementSet version 2.0 as the foundational standard, and all subsequent analysis and discussion follow this version specification.

1.2.1 Subtable Structure The table structure used by CASA for MeasurementSet is shown in Table 1 . Each MS file must have a main (MAIN) table containing numerous data columns and keys to various subtables. Each subtable appears at most once. Subtables are stored as keywords of the MS, with all defined subtables listed below. Optional subtables are shown in italics and parentheses. In practical applications, non-optional subtables must be created during data generation, though some subtables may be empty.

Table 1. Table structure for MS V2 version

The structure is significantly more complex than FITS files. In practice, according to the specific observation data requirements of different radio telescopes, corresponding subtables are generated and relevant data are stored to form a complete MS file directory tree structure. In principle, all non-optional tables must be created. Due to space limitations, we analyze the relevant field requirements and data types of the MS format using the main table structure in the following section; the structures of other tables can be examined directly in the MS technical specification.

1.2.2 MAIN Table: Data, Coordinates and Flags The main table (MAIN TABLE) is a required component of an MS file. In terms of data storage, the MS format is broadly similar to FITS, needing to store data types including integers, floats, doubles, strings, etc.

Similar to FITS header definitions, each table in an MS file has a corresponding field design concept, but it is more complex. MS field designs fall into three categories: 1. **Keywords**: Such as MS_{VERSION}, which identifies which version specification the MS file follows. 2. **Keys**: Equivalent to primary keys in relational databases, used to associate with subtables. For example,

TIME provides the observation timestamp. These values must be defined and written strictly according to MS format requirements. 3. **Non-key attributes**: Certain important parameters or attributes defined according to actual needs.

The storage of other tables is completely analogous to the main table, with their own specified reserved words, keys, and parameters. Successfully generating the MS format essentially requires clarifying the attributes and format requirements of each field, writing correct values, and providing correct units.

1.3 MS File Storage Structure

Unlike typical relational databases that consist of only database files and index files, the MS format employs a notably different approach using a multi-level directory and multi-file storage method. Each table is stored in CASA table format, meaning a single table comprises multiple files, and the entire MS file is not a single file but a directory tree.

The MS file directory structure can be viewed as composed of multiple levels. Generally, the main table resides in the first-level directory, while various subtables are located in the second-level directory. Each level of directory contains files such as table.info, table.f0, table.f1, etc., which store the actual data location information.

II. MS Generation Method Based on Python-Casacore

As is well known, since the evolution from AIPS++ to CASA, the CASA software development has adopted a multi-language programming approach, with its main code derived from C/C++ and the primary user interface and invocation layers implemented almost entirely in Python. The C/C++ components are consolidated into the casacore software package, known as the CASA core library. Casacore is currently the most comprehensive radio astronomical data processing software package and the only one that implements MS file read and write operations.

To enable calling casacore functions from Python, python-casacore was developed as a wrapper for the core function library. Although documentation is available, detailed instructions on implementing MS format writing are lacking. Therefore, through extensive experimentation and verification, we analyzed the basic function calls in python-casacore and examined the functions and parameter types for operating on data tables. Building upon this foundation, we further encapsulated a complete MS format output object and integrated the code into ARL, enabling comprehensive MS file generation through simple calls to this object instance.

2.1 Table Operations in Python-casacore

Python-casacore is a Python wrapper interface for casacore. For MS file data operations, it primarily provides the functions shown in Table 2 .

Table 2. The MS-related functions of python-casacore

2.2 Subtable Generation

Using the methods in Table 2, a complete subtable can be generated. Table 3 presents a program flowchart for generating an MS file instance, illustrating how to apply python-casacore to create corresponding tables and populate them with data.

Table 3. The demonstration codes for MS generation

III. MS File Generation in ARL

The Algorithm Reference Library (ARL) is a radio interferometric data processing algorithm validation library developed by radio astronomer Tim Cornwell to provide algorithm verification for subsequent SKA data processing. During our involvement in the bridging phase, we developed an MS file output module based on ARL to save simulation results and enable data sharing with other common astronomical data processing software for validating ARL outputs. The final encapsulated MS file class is WriteMS, with its class diagram shown in Figure 1 [Figure 1: see original paper].

In practical applications, users simply need to import the software package and pass in the corresponding visibility function data to generate a complete MS file. A complete implementation code is provided in ARL (test_{{{export}}*{{ms}}}{arl}.py). Figure 2 [Figure 2: see original paper] illustrates the MS file writing and generation process using a program flowchart for a simulated MS file generation example.

IV. Results Testing and Verification

To verify the correctness of MS file writing, the simplest method is to read original MS format observation data, rewrite the data using our method to generate a new MS format file, process the newly generated MS data with CASA software for imaging, and compare the imaging results with the original observation results. Since CASA software performs numerous validation operations when reading MS data, successful processing by CASA indicates that the various subtables and fields in the MS file are reasonable. Although values in some fields may differ from actual conditions, this does not affect imaging processing. The specific operations are shown in Table 4 .

Table 4. The commands for imaging in CASA

Figure 3 [Figure 3: see original paper] shows the dirty image obtained by CASA reading the MS file and performing imaging (right panel) alongside the image used for the simulated observation (left panel). The comparison reveals that the imaging results based on the MS file generated in this study are consistent with the actual image, thereby verifying the correctness of the MS file.

V. Conclusion

Although the MeasurementSet file specification was established relatively early, it has seen limited application in China's radio astronomy field. This is partly because MS files occupy considerable space, and partly because generating MS files has long depended on the casacore underlying software package, making development difficult. Therefore, this paper systematically investigates and discusses the definition, structure, field design of MS files, and data writing using Python-casacore, all in the context of SKA engineering construction needs. The final experiments demonstrate the correctness of the generated MS file content. All relevant code has been integrated into the ARL software, with all source code available for download at https://github.com/SKA-ScienceDataProcessor/algorithm-reference-library. Overall, this work has played a crucial role in ARL development, providing guarantees for subsequent SKA data simulation and file storage, and serving as a valuable reference for MS file generation for other radio astronomical data.

VI. References

[1] Wu X. Probing the epoch of reionization with 21CMA: status and prospects[C]. Bulletin of the American Astronomical Society, 2009: 474.
[2] 南仁东. 500m 球反射面射电望远镜 FAST[J]. 中国科学 G 辑: 物理学、力学、天文学, 2005, (05):
[3] 陈学雷. 暗能量的射电探测——天籁计划简介 [J]. 中国科学: 物理学力学天文学, 2011, 41(12): 1358-1366.
[4] Yan Y, Zhang J, Huang G. On the Chinese spectral radioheliograph (CSRH) project in cm-and dm-wave range[C]. 2004 Asia-Pacific Radio Science Conference, 2004. Proceedings., 2004: 391-392.
[5] Offringa A, Mckinley B, Hurley-Walker N, et al. WSCLEAN: an implementation of a fast, generic wide-field imager for radio astronomy[J]. Monthly Notices of the Royal Astronomical Society, 2014, 444(1): 606-619.
[6] Hamaker J, Bregman J, Sault R. Understanding radio polarimetry. I. Mathematical foundations[J]. Astronomy and Astrophysics Supplement Series, 1996, 117(1): 137-147.
[7] Petry D. Analysing ALMA data with CASA[J]. arXiv preprint arXiv:1201.3454, 2012.
[8] Kemball A, Wieringa M. MeasurementSet definition version 2.0[J]. URL: http://casa.nrao.edu/Memos/229.html, 2000.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.