Formal Representation of the MeasurementSet

F. Viallefond, LERMA, UMR8112, Observatoire de Paris, 61 av. de l'Observatoire 75014 Paris France

Abstract

Models have been defined by the community for the data generated in radio observatories. These models have evolved with time to match the evolution of the instruments. We demonstrate that the MeasurementSet as specified in 2000 remains a solid conceptual basis to start from to develop models for the new generation instruments. From our analysis has emerged a theory allowing us to formalize models. This theory appears useful not only within the context of the on-going major evolution of the radio-astronomy instrumentation but also in other domains. It relies on the topology, the axioms of the theory of groups and algebraic geometry. On the practical side this theory can be used to assess coherence when elaborating concepts to describe physical systems.

1. Introduction

About 20 years ago radio-astronomical institutes had federated their efforts to develop a common platform for the processing of their observational data, especially the calibration and the imaging. This was the AIPS++ project which is build on two pillars, the MeasurementSet (MS) and the MeasurementEquation (MEQ). In contrast to the later which has a solid basis thanks to the formalism developed by [1] the former is still lacking firm foundations. One of the corollary is the lack of a calibration data model (CDM) that we would like to have to bridge the MS with the MEQ and of a grand unification to include the structures involved in imaging part, all these things being connected each other. As a side effect the MS, so far merely used through a spreadsheet, has never been implemented by mean of a formal structure associated to an algebra.

This paper reports about a formalism elaborated to describe rigorously the MS. Because this work uses mathematics with a high level of abstraction it will be described elsewhere. In the following we merely provide some hints by giving the path which has been followed to obtain to our main result, the existence of a *theory*. Although this result has implications far beyond the scope restricted to the MS use-case it is worth mentioning that the exercise of describing concisely the physical systems that we use in radio-astronomy has been a major source of inspiration helping to reveal this theory. The MS provides a very concrete application which can be used as an example to explain how this theory could be applied in other domains. Restricting to our domain, we address in this paper the question of the future of the MS like models. It can be demonstrated that we are on the right track. Therefore we can safely conclude that the MS with necessary well defined evolutions forms a very solid basis to start with for the definition of a data model for the new generation radio-telescopes as those anticipated for SKA. With this theory one may envision multiple practical applications. For example we may envision a tool with computer proof assistance to assess the cohesion of our models when defining concepts to build physical systems. As an other example we are relatively confident that the formalism resulting from this theory will give us the possibility to fill this gap between the MS and the MEQ. In any cases this theory gives a rich conceptual framework with significant potentials to contribute in several domains. Looking forward for an unification we do not exclude that our approach could also be useful to contribute to the on-going evolution of our way of imaging from the measurements.

2. Context of the work.

Our community has defined models to represent the data produced by radio-telescopes. These are used during the observations for telescope calibrations and offline by software packages for data reduction, calibration and imaging. A major step has been the definition of the MS in [2] for offline processing. Observatories have developed their models for their own instruments, *e.g.* ALMA with its Alma Science Dada Model (ASDM). Although this ASDM reuses most of the concepts present in the MS it has distinct features. One of the most striking is the introduction of the concept of configuration, a necessary evolution as the telescope hardwares offer extreme flexibility with numerous possible configurations. The ASDM has also numerous enumerations; this can be seen as a domain specific vocabulary. Reused for the EVLA the acronym is simply 'SDM'.

978-1-4673-5225-3/14/$31.00

Boosted by the SKA project the instrumentation is evolving rapidly. Not only the data flow will increase dramatically as already seen with precursors such as LOFAR but the description of the data will also become much more complex. These are several reasons for that, *e.g.* the fact that high dynamic range will be required to achieve the theoretical sensitivity implies to describe the hardware with much more details in comparison to what we are used to so far. These new instruments will also have more automated data processing up-stream; this will produce more meta-data *e.g.* triggers to send directives down-stream in the work-flow. In spite of the fact that the instrumentation will cover a broad range of different concepts, observational procedures and kinds of scientific projects a number of software components will remain of common interest in that diversity. In any cases our way of managing and processing data has to evolve very significantly. Getting high performances is critical for all observatories, not only those of the future but also those with development plans for existing telescopes.

To identify and evaluate the concepts used at the foundation of the MS and its potential evolution we *a)* account for lessons (the good and bad things) that we learned through the development of the SDM, *b)* use our R&D work, a very concrete and practical activity within the environment of EMBRACE, a project to prototype and evaluate the concept of dense aperture arrays (DAPA) and *c)* use mathematics to develop all the theoretical aspects.

This exercise is performed with three keywords in mind, *expressiveness*, *efficiency* and *robustness*. Having the software written with expressiveness is useful all along the life cycle of the codes making them easier to maintain and evolve, this being done by peoples located all over the world with different expertise. Efficiency is obviously very important; we must understand the performances measured on realizations. Robustness is required as we may have highly complex software components and we must assess the exactness of the results. To have these three requirements we use generic programming techniques, therefore a strongly typed language. This approach which requires a high level of abstraction leads naturally to formalizations, hence mathematics.

3. Theoretical aspects.

This section gives an overview of the path which we have been using to find and elaborate this theory. From now we will consider our model which is a generalization to cover the use-case of DAPAs. This model was constructed in 2010, mostly intuitively, by combining numerous features of the MS as specified in 2000 and features that we had developed for the ASDM. This model was defined and implemented *before* making the connection with the mathematics. Because of this generalization and of the way we did its implementation this model is more generic. An analysis of this model shows that there are very similar structures to describe not only the concepts of SpectralWindow (the frequency axis) and Antenna (the aperture axis) but also our way to describe the time axis. The former originates from the MS and the later from the SDM. Most striking was what was done to cover the DAPA use-case this allowing us to discover this similitude. These three concepts form the primary axes of the model.

An important step in this analysis has been when we realized that it was possible to describe the high level structure of the whole model with a *directed graph* which has a well defined geometrical signature. Then we realized that it was possible to reuse this graph to describe things at various levels, *e.g.* our way to describe our Antenna concept. These graphs are found using reasonings to describe concepts with our human language giving a logical meaning. Although this graph is planar it is also thought in 3D by folding parts of it to reveal triples which appears to be relations carrying meanings. Formally it is a simplicial 2-complex. The discovery of this graph was not completely be mere coincidence because in the mean time we had made a connection with a branch of mathematics which gives us the proper language to explain how we did our generic implementation of the algebra of the physical quantities (the MS uses the concept of measurement which derives from physical quantities). Our generic are based on topological spaces with adjunction of functors. Taking the arrows in our graph for functors between identified topological spaces this directed graph corresponds to a diagram.

The next step was when we realized that we can also use this graph completely outside our radio-astronomy domain, in particular in computer sciences to describe the component model in the grammar of the XMLSchema language but also in the domain of mathematical physics. This suggests the existence of a theory these different use-cases being applications of that theory to model systems in different domains.

The theory was found by mean of an action group. It is a method used in algebra and geometry to describe objects using groups of symmetries. We obtain this diagram by invoking Hamiltonians vector fields and the construction of

chain complexes which are finite sequences. This result being independent of the domain of applications that theory is effectively obtained. The concept of chain complexes is at the foundation of the homological algebra. Indeed our diagram contains both a topological structure and an algebraic structure.

4. Applications and practical considerations.

This theory allows us to go deeper in the understanding of the structure of the model of the radio-astronomy. Indeed the definition of the MS in 2000 and the prototyping activity to investigate its evolutions were done ignoring the existence of that theory. As a matter of fact the intuitions and reasonings used to elaborate these structures have worked remarkably well! To illustrate this let us take one example, the location where we invoke the polarization in the MS. It is an attribute in the Feed table. This theory tells us that this is precisely where it has to be! What is this location? The theory shows that the key section of this table corresponds to a homotopy category chain complex and that the polarization must be at the end of an exact sequence, the one where there is a *choice* between two Hamiltonian vector fields mirror images of each other. This example highlights the fact that the geometry plays a very important role. The corollary is that we can use representations with matrices. Quite naturally these results are highly suggestive to bridge the MS and MEQ using this formalism.

Throughout this work we found significant connections to the notion of type. Our diagram reveals an internal structure which is a tensor. In some way we have extended the notion of type to the case of an object which is defined by a compound of parts independent from each other but cohesively bound. It is instructive to look how this idea is reflected in our model. It comes from the kind of model structures that we use to represent the axes in our model and the fact that our physical systems involve properties in the direct and the image domain, *e.g.* the concepts of aperture and beam.

Some anticipated applications of this theory have been introduced in sect. 1. On the software side we think that there is a way to implement our models more generically compared to what we did in 2010. The codes written using generic programming techniques are not always easy to read! It must be understood that the expressiveness is at the application level, that is in the codes invoking the generic templates. Robustness comes whenever there is algebraic expressions between types.

Remarks about the practical side:

We do not expect all the users to know much about generic programming. Expressiveness means declaring variables with types using terms from a domain specific vocabulary. These terms are defined in what we call a *profile*. The user has to set up a profile to describe his/her physical system. Profiles must be valid instance documents against a schema. This schema assess that the structure which is defined is coherent. This is what we did for our prototype of 2010. We plan to continue with this strategy with the following evolution: *a)* the schema will result from this theory and *b)* the software components which process these profiles will be modified accordingly. Most of the generated code will be template specializations and assignments to the domain specific language given by the user.

Nor do we expect the users to know about the actual terminology used to explain the theory in mathematical terms. The schema provides an assistance when editing profiles. This schema may be quite sophisticated. For example for in our prototype of 2010 it is possible to specify in the profile structured use-cases, data-stream descriptions etc. These informations are processed to produce a user-oriented documentation filtered for a given use-case, *e.g.* a DAPA system. This allow the user to focus only on the things which are relevant to his/her domain. This our way manage models for heterogeneous systems as expected, the coexistence of *e.g.* DAPA and classic dishes.

5. References.

1. Hamaker, J., Bregman, J.D., Sault, R.J. 1996, A&A **117**, 137

2. Kemball, A., Wiringa, M. 2000, http://casa.nrao.edu/Memos/229.html