The 5^{th} IVS General Meeting Proceedings, 2008, p. 143–152

IVS Working Group 4 on VLBI Data Structures

John Gipson

NVI, Inc./NASA Goddard Space Flight Center, USA

Abstract. In 2007 the IVS Directing Board established IVS Working Group IV on VLBI Data Structures. This note discusses the current VLBI data format, goals for a new format, the history and formation of the Working Group, and a timeline for the development of a new VLBI data format.

1. Introduction

At the Sep. 15, 2007 IVS Directing Board Meeting I proposed establishing a "Working Group on VLBI Data Structures". The thrust of the presentation was that, although the VLBI database system has served us very well these last 30 years, it is time for a new data structure that is more modern, flexible and extensible. This proposal was unanimously accepted, and the board established IVS WG4 (Working Group IV) with myself as chair.

One motivation for changing the data structure now is the advent of VLBI2010. The first VLBI2010 antennas are due to come on line in 2010, and the number is expected to increase rapidly. Because we want to tie these stations into our current geodetic network, these stations will have the capability of observing at S/X. However, 'native' VLBI2010 observing will be very different than our current observing in many ways. The receiving hardware will be broadband and observe over a much wider frequency range. It will likely record over many bands, as opposed to our current 2, and even the concept of "band" may no longer be meaningful. The system will record many more bits, which will result in greatly increased sensitivity. This will allow the use of small, fast antennas, which, combined with the increased sensitivity, will enable them to participate in on the order of 100 scans/hour instead of the current 10-15. Current operational sessions such as the R1s and R4s typically involve around 7 stations and make about produce $\simeq$ 3000 observations per session. We expect that the number of stations participating in a typical VLBI2010 session will increase to 16, 20, 32 or even more. Although the current RDV sessions may involve up to 20 stations, they are the exception rather than the rule. Simulation studies indicate that a 24-h VLBI2010 sessions with 16 stations may yield 100000–200000 observations, and larger networks would have

even more observations. In short, life will be much different when VLBI2010 is fully operational.

In this note I begin by reviewing the history of the VLBI database format. I then discuss some of the features and limitations of the current structure. This is used to motivate the desirable features of a new data structure. Following this I discuss the history, formation, composition and goals of WG4. I summarize the first meeting of WG4, and the next steps. Lastly I offer some concluding remarks.

2. Brief History of VLBI Database Format

The VLBI database format and associated software was developed in the mid-1970s, at the very beginning of geodetic VLBI. It was designed at NASA's Goddard Space Flight Center (GSFC) for use with the *calc/solve* analysis package, and is closely tied to this package. It is currently the IVS standard for archiving and distributing geodetic VLBI sessions. It is exceptional for anything in the software world to still be in use after 30 years. On the one hand, this is a testament to the design and robustness of the VLBI database format, which has been pushed to do things well beyond the original specifications. On the other hand, the VLBI database is a product of when it was designed, and suffers from this. The world has changed in the last 30 years: the VLBI technique has continued to evolve and mature; the speed and power of computer hardware has increased by almost 4 orders of magnitude; software has become more flexible and powerful. There is no doubt that if we were designing the database system today it would be much different.

3. Features of Current Database Format

In this section I review some of the features of the current VLBI database format. I want to emphasis that this is not done to disparage the current system, which has served us well. Rather, it is to see how it could be improved.

Provenance. A very positive feature of the database structure is that it encourages analysts to keep a history of what was done and by whom.

Redundancy due to baseline orientation. The VLBI database system was developed when VLBI sessions involved single baselines observing a single-band. Because of this the database is baseline observation oriented. In order to process and analyze an observation, a lot of information is needed, and most of this information is stored for each observation. However much of this information is the same for other observations within a scan. Since there are $(N-1) \times N/2$ observations in a scan, this leads to a tremendous amount of redundancy. Tabl. 1 gives examples of three kinds of data associated with an observation. **Baseline** dependent data varies from observation to observation within a scan. **Station & Scan** dependent data is common to all observations involving a given station within a scan. This only needs to be stored once per station per scan, instead of the current $(N-1)$ times. **Scan** dependent data is

the same for all observations within a scan, and needs to only be stored once, instead of the current $(N-1) \times N/2$ times. Tabl. 2 calculates the amount of redundancy for the different data types as a function of the number of stations in the scan. The effect is negligible for small networks but grows rapidly.

Table 1. Types of VLBI data associated with an observation

Baseline Dependent	Station & Scan Dependent	Scan Dependent
Delay	Met Data	Source
Rate	Elevation	Epoch
Ambiguity	Phase Calibration	EOP
Ionosphere Calibration	Cable Calibration	etc.
etc.	etc.	

Table 2. Data redundancy and storage efficiency for VLBI data types

Scan Size		Data Redundancy			Storage Efficiency		
Stations	Baselines	BL	Stat	Scan	BL	Stat	Scan
2	1	1	1	1	100%	100%	100%
3	3	1	2	3	100%	50%	33%
5	10	1	4	10	100%	25%	10%
10	45	1	9	45	100%	11%	2%
15	105	1	14	105	100%	7%	1%
20	190	1	19	190	100%	5%	1%
32	496	1	31	496	100%	3%	0%

Custom Format. Databases are written in a special format, and require special software—the database handler—to read and modify them. This was not an issue when everyone used the GSFC developed *calc/solve* analysis package, but became one with the development of alternative analysis packages. This software was written in FORTRAN in the 1970s and 1980s. It incorporates limitations of FORTRAN at that time. For example, the earlier code does not recognize character strings, and still uses Holleriths.

Speed of Retrieving Data. The database handler was written before the advent of modern database systems. It is slow in retrieving data, and can take many minutes to retrieve all of the data in a VLBI database. Because of this, institutions, including GSFC, have developed their own internal formats for processing and storing VLBI data.

Irrelevant/Obsolete Data. The VLBI database contains data which are no longer used. Some of this is due to computer hardware issues present at the time the database was designed, when numerical calculations were very costly. For example, the databases contain the numerical value for π.

Data used only by *calc/solve*. The database contains items, e.g., partial derivatives, which are computed by *calc* and stored for later use by *solve*. Other analysis packages do not use these or compute them "on-the-fly". Some of these items are not even used by *solve*. When *solve* was the only VLBI analysis package, this was not an issue. In principle, the IVS standard for archiving and transmitting data should not give preferential treatment to one analysis package.

Lack of Flexibility. The only people who can add new kinds of data to the database are those who have installed the database handler. In practice, this means individuals or institutions who have installed the *calc/solve* analysis system. This is an additional obstacle in improving the analysis of VLBI data.

Incomplete Data. When the database was originally designed, it was envisioned that it would contain all of the data necessary to process a VLBI session. However, there are currently many data items used in the analysis of VLBI sessions that are not contained in the database. These include EOP, pressure loading calibration, external mapping functions (e.g., VMF), etc.

Coupling of observables and models. In some sense, the database treats all data as equal. The observables, which never change, are on the same footing as geophysical models. They all live in the database together. One result of this is that as geophysical models change, the databases are updated. Although this is not required in principle, it is necessary at GSFC in order to test these models. This makes testing new models at GSFC cumbersome and time consuming, and results in changes in the version number of database when nothing fundamental has changed.

No Access to Primitive Data. The lowest level of data stored in the databases are the various kinds of delay output by the correlator program *fringe*. Although this is sufficient for most people most of the time, occasionally there are problems with the *fringe* output. The origin of these problems can sometimes be determined by looking at the raw-correlator output, which serves as input to *fringe* The database contains no natural "hooks" to this data. For experts it may be useful to have the ability to re-fringe the data using different assumptions, or, perhaps different algorithms.

4. Goals for the New Format

At a minimum, any new data format must be able to store the data currently required to process VLBI sessions. It should also handle the anticipated needs of VLBI2010 and beyond. Without these, there is no point at all in designing a new system. The current database system could be modified to handle the needs of VLBI2010 with considerable effort. This would still leave us with the current problems.

Based on the discussion of the previous section, following are some initial goals for a new VLBI data format.

1. Provenance. Analysts should be able to determine where the data came from and what happened to it.

2. Compact. The database structure should minimize redundancy.
3. Accessible. Users should be able to easily access the data without the need of custom software. Currently we do have such a format—NGS cards—but this suffers from other problems.
4. Cross platform/OS/language support. The same structure should be accessible by programs written in different languages running on different computers using different operating systems.
5. Speed. Users should be able to add, modify and retrieve data quickly.
6. Extensible. It should be possible to add new data types, e.g. source-maps, antenna temperatures, system gain information. Some of this information may not be currently useful, but may become so in the future.
7. Completeness. The data structure should include all data necessary for processing VLBI data. Analysts should be able to redo the entire analysis from start to finish.
8. Separable. Analysts should be able to retrieve just the data they are interested in.
9. Decoupling. There should be a clear distinction between fundamental observables (e.g., delay) and other data items.
10. Different levels of abstraction. There are many different kinds of users of VLBI data. The new structure should serve all of these equally. Most users may be interested only in the final delay. Experts may want access to data at a more primitive level.

5. History and Formation of the Working Group

At the 2007 IVS Analysis Workshop held in Vienna there was extended discussion about the need to revamp the VLBI data structure. In May I circulated a draft proposal to the IVS Directing Board proposing the formation of a Working Group for this purpose, and offered to chair it. At the request of the IVS Directing Board, I made a presentation at September 2007 Board meeting. The IVS Board unanimously approved the formation of IVS WG4.

Any change to the VLBI data format affects everyone in the VLBI community. Therefore, it is important that the working group have representatives from a broad cross-section of the IVS community. Tabl. 3 lists the current members of WG4 together with their affiliation. The initial membership was arrived at in consultation with the IVS Directing Board. On the one hand, we wanted to ensure that all points of view were represented. On the other hand, we wanted to make sure that the size did not make WG4 unwieldy. The current composition and size of WG4 is a reasonable compromise between these two goals. My initial request for participation in WG4 was enthusiastic: everyone I contacted agreed to participate with the exception of an individual who declined because of retirement.

Table 3. Membership in Working Group IV

Chair	John Gipson
Analysis Center Director	Axel Nothnagel
Haystack/ Correlator Representatives	Roger Cappalo Colin Lonsdale
GSFC/Calc/Solve	David Gordon
JPL/Modest	Chris Jacobs
Occam	Oleg Titov Volker Tesmer Johannes Boehm
Main Astronomical Observatory/ Steelbreeze Observatorie de Paris/PIVEX NICT	Sergei Bolotin Anne-Marie Gontier Thomas Hobiger Hiroshi Takiguchi

Communication is crucial to the success of WG4. This includes both communication between the members, and communication between WG4 and the broader IVS community. Following the lead of the VLBI2010 committee, I anticipate that much of our work will be done via email and teleconferences. Our email discussions are publicly available on the IVS web-site. In addition, we plan on having at least two face-to-face meetings each year. For the convenience of participants, these will occur as splinter meetings at the AGU, EGU and IVS General Meetings. These meetings are primarily for WG4 members, but other interested parties are encouraged and welcome to attend. We will also make regular presentations at the IVS General Meeting and other conferences as appropriate. Our goal is to make sure that everyone knows what we are doing, and that there will be no unpleasant surprises.

In addition, I or another member of WG4 will update the IVS Directing Board at each of their board meetings. This will keep the Board apprised of

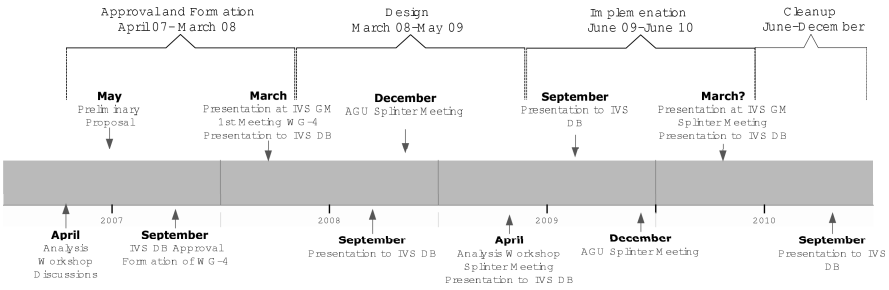

Figure 1. Past and future of working IVS WG4

what we are doing, and also serves as an opportunity for feedback. The current plan is that WG4 will be dissolved at the end of 2010.

Fig. 1 is a timeline of the past and future of IVS WG4 from its inception through 2010. Marked on the timeline are the proposed meetings of WG4, regular reports to the IVS Directing Board, and the various phases in the life of the WG4.

Table 4. Partial list of data to be included in new format

VLBI data	
Correlator output	
Original fringed data	Refringed data (if any)
Editing criteria	
Group delay ambiguity	Phase delay ambiguity
Fully calibrated and ambiguity resolved group delay	Fully calibrated and ambiguity resolved phase delay
VLBI calibrations	
Raw Phase calibration	Modified Phase calibration
Raw Cable calibration	Modified Cable calibration
Other raw calibrations	Other modified calibrations
Associated data	
Raw met data	Calibrated met data
Location of met sensors	
Physical antenna temperatures	Antenna height measurements
System temperatures	Gain measurements
Pointing measurements	Source maps
Geophysical models/effects	
Pressure loading	Ocean loading
Calculated mapping functions e.g., VMF	Slant path delay
Earth orientation parameters	
Miscellaneous files	
Schedule file	Experiment notes
Log files	Emails
Correlator summary	

6. Summary of Initial Meeting of Working Group

WG4 held its first meeting at the IVS General Meeting in St. Petersburg. This meeting was open to the general IVS community. Roughly 25 scientists attended: 10 WG4 members, and 15 others. This meeting was held after a long day of proceedings. The number of participants and the lively discussion that ensued is strong evidence of the interest in this subject. By the end of the

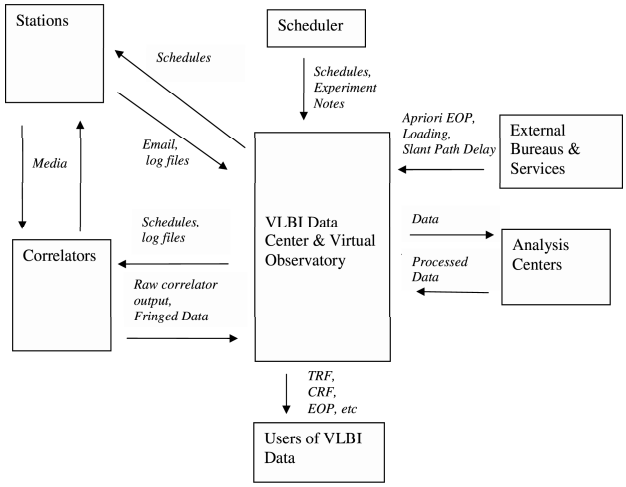

Figure 2. Under the proposed scheme a "virtual observatory" serves as the central depository of all VLBI related data

meeting we had arrived at the following consensus.

1. Users should be able to trace the origin and the history of all the data: where it came from; when it was processed; who processed it; what procedures they used, etc.
2. All VLBI data should be easily accessible from the web. This includes the whole data-flow of VLBI, from scheduling the session, through correlation and analysis. Tabl. 4 provides a rough draft of data that would be accessible.
3. VLBI data should be available at different levels of processing and abstraction. This ranges from raw correlator output, which rarely changes, to final edited and calibrated data. Experts should be able to access the data at a low level, but other users of VLBI data, who lack interest or expertise, should be able to access edited and calibrated data.
4. Other data used in processing and analysis of VLBI data should also be available at different stages of processing and calibration. For example, raw met data should be available, as well as smoothed and calibrated met data. This would allow research into alternative methods of calibration and smoothing.
5. The format should also include data which is externally derived, but used in VLBI analysis. Examples of this include pressure loading corrections, ocean loading corrections, antenna tilt corrections, etc.
6. Stations, Correlators, Analysis Centers and others would all contribute data, and all be viewed on an equal footing.

7. Analysis centers should be be encouraged to post the results of their processing. Examples include: editing criteria used; ambiguity resolution; calibration information; source maps or source corrections; etc. This would make it possible for others to use these results, and would also help ensure transparency in how data is analyzed.

Tabl. 4 is a first cut at the kind of information that would be accessible. As time goes on, other kinds of data will surely be added. In some sense, the above considerations imply a combination and extension of the current VLBI databases, the information contained on the IVS session web-pages, and lots more information. The idea of a "Virtual Observatory" for VLBI data arose as an outgrowth of these discussions. Fig. 2 is a schematic illustration showing how different institutions would interact with this.

The desires expressed in the first meeting were much more ambitious than I originally envisioned, and it is unlikely that all of these goals will be accomplished in the limited span of WG4.

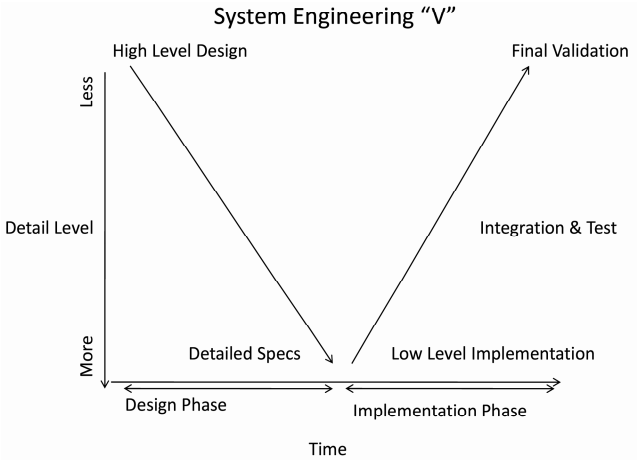

Figure 3. For optimal results, you should spend as much time on design as implementation

7. Next Steps

Fig. 3 is commonly used in system engineering to describe the steps involved in developing and implementing new systems. This is a useful paradigm for WG4. The system engineering process is divided into two parts: design and implementation. Design starts at a high level of abstraction and ends at a low level. Initially you define what your system is supposed to do, and then you start discussing how to do this in more and more detail. Implementation does just the opposite: you start at the bottom, building small parts of the system,

and work your way up, integrating the small parts together. Periodically you stop to test and validate what you are doing.

A natural tendency is to jump into implementation too soon. You think you know what you want, and start to build it. Experience has shown that the more time spent on specification and design, the less time you will spend fixing problems. Ideally, you should spend about as much time in the design part of the process as in the implementation part.

Based on the above, the first goal of WG4 will be to have a complete specification by mid-2009. Progress to date will be presented at the European VLBI meeting in April 2009. This will also provide an opportunity for IVS members to give feedback. Following this, we will begin the process of implementation. Our goal will be to have this completed by the end of 2010. We will give another status report at the IVS General Meeting, which will provide a chance to make last minute course corrections.

We are currently in the very early design stages. As part of this effort we will see what other groups with similar data handling issues have done. We anticipate that we will be able to borrow from and build on the work of others. For example, there are currently several excellent public domain structures to deal with various kinds of scientific data, such as NetCDF, FITS, HDF etc. There are interfaces to these data structures on a variety of platforms and from a variety of computer languages. Each of these has a large community of users, and each has and a large library of freely available utilities for extracting and handling data. We will study these formats in detail to see how well they would work for VLBI data and what the trade-offs are.

8. Conclusion

The goals of IVS Working Group IV are very ambitious. We can only achieve these goals if all of the members contribute to this effort. The Working Group will endeavor to operate in an open environment. We will make regular presentations to inform the community of our progress, and solicit input from all interested parties.