The VLBI Data Interchange Format (VDIF): 13 Years On

Mars Buttfield-Addison [1]

[1] *University of Tasmania, Churchill Ave, Hobart, Tasmania 7005*

Summary: VLBI relies on incorporation of data from multiple sensors that may vary dramatically in hardware and behaviour. Data consistency is critical, so the VLBI Data Interchange Format (VDIF) standard was developed—a format that could be used for VLBI data, both in transit and at rest. VDIF has achieved widespread adoption, but compatibility issues remain. Anecdotally, issues such as undocumented systemic errors, format non-compliance, and institution-specific workarounds appear commonplace. This means that when attempting to analyse VDIF data, researchers may be lacking the contextual knowledge required to accurately interpret that data. To ensure the integrity of future research, internal issues and workarounds must be exposed, documented, versioned, and archived in perpetuity. A survey was conducted among institutions using the VDIF standard, asking respondents about their experiences as producers and consumers of VDIF data—and requested a sample of data that each have produced. Several key issues were identified, among even well-established producers of VDIF data.

Keywords: VDIF, data formats, computing, compliance, VLBI.

Introduction

Very Long Baseline Interferometry (VLBI) is a method used in astronomy where simultaneous observations with multiple radio telescopes separated by great distances are combined to emulate a single, much larger telescope. The sensitivity gained from VLBI has allowed for high resolution imaging of sources in deep space, tracking of distant spacecraft, and new levels of precision in the fields of astrometry and geodesy [1, 2]. But the core of the method relies on precise, efficient, and reproducible incorporation of data from multiple sources.

Prior to 2009, the format of VLBI data produced by an individual institution was primarily determined by the data capture hardware itself—for example, the popular Mark5B recorder and associated data format [3]. Following this, a new format was devised by an expert task force with input from the broader VLBI community, to unify the field under an international standard: the VLBI Data Interchange Format (VDIF) [4]. Its design prioritised ease of data transmission, processing, and storage, and made accommodations for the operating conditions of VLBI systems—such as allowing for fast, out-of-order processing of high-volume data from multiple sensors.

The format was well-received by the community and, in the thirteen years since its ratification, has seen the widespread adoption necessary to achieve the status of industry standard [5]. By this metric, the VDIF project and specification is a resounding success. But adoption rate is not the only goal: an effective data or software standard must also be simple to adopt, comply with, verify compliance with, and adapt to specific user needs. This study works to gauge these factors of the VDIF format, with a focus on identifying persistent format compliance and usability issues among its users.

The VDIF Format

VDIF data is structured in often large binary data files with a strict internal structure and file naming scheme [6]. Each file is made up of a series of frames, each of which contains a structured **metadata header** containing details about the recording properties and conditions, and an **encoded data array** which should be unpacked and decoded as per the values in the header. This design is highly efficient for data capacity, and allows the data represented in each frame to be fully understood with minimal additional information[1] but means that minor errors or omissions in the header of a frame can lead to dramatic differences in decoded data.

To allow for integration of multiple data capture devices or modes—such as in multi-telescope arrays or interferometers—VDIF files support multiple **data threads**, indicated by a thread ID value in each frame header. A backend recorder will typically capture individual frames as they are received from each source, leading to interleaved threads in the final recording that may be *out-of-order* (e.g. subsequent frames on disk from thread 0 and thread 1 may have dramatically different timestamps) or *skewed* (e.g. subsequent frames from a particular thread on disk may be separated by a large number of frames from other threads). This complicates the matter of processing VDIF data, as though the specification may leave such phenomena unbounded, each realised implementation of the format will be inevitably limited by technical aspects such as system memory. Because these may vary between implementations, this introduces further potential inconsistency in decoded data or processing behaviour from the same VDIF source.

Full details of the VDIF specification can be found at https://vlbi.org/vlbi-standards/vdif/.

Identifying VDIF Non-Compliance

To identify issues faced by producers and consumers of VDIF data in their day-to-day work, a survey was conducted of institutions that either operate VLBI components or act as correlator sites for VLBI experiments. Recruitment was primarily done via mailing lists and direct email contact, targeting technical staff in roles such as instrumentation or software development within such organisations. Prior to being presented with the web-based survey, potential participants were asked to confirm their employment at the given institution—by providing an email address from the relevant domain—and assert their suitability to provide technical details across key criteria.

So far, the survey has received 25 responses, representing a spread of geographical locations and both public and private institutions. The institutions represented can be seen in Table 1.

Table 1: Respondents to the VDIF survey, by institution of employment.

Institution or Observatory	Primary Location
Atacama Large Millimeter Array	Antofagasta, Chile
Auckland University of Technology	Auckland, New Zealand
C.A. Muller Radio Astronomy Station	Dwingeloo, Netherlands
Canberra Deep Space Communications Centre (NASA)	Australian Capital Territory, Australia
EISCAT Scientific Association	Lappland, Sweden
Geodetic Earth Observatory (Kartverket)	Svalbard, Norway

[1] In the default VDIF implementation, data sample rate must be inferred through examination of multiple frames.

Green Bank Observatory	West Virginia, USA
Hartebeesthoek Radio Astronomy Observatory	Gauteng, South Africa
Haystack Observatory (MIT)	Massachusetts, USA
Istituto di Radioastronomia	Emilia-Romagna, Italy
Jodrell Bank Observatory	Cheshire, England
Max Planck Institute for Radio Astronomy	Bonn, Germany
Metsähovi Radio Observatory	Kirkkonummi, Finland
National Astronomical Observatory of Japan	Tokyo, Japan
Onsala Space Observatory	Halland, Sweden
Parkes Observatory (CSIRO)	New South Wales, Australia
Shanghai Astronomical Observatory	Shanghai, China
Smithsonian Astrophysical Observatory	Massachusetts, USA
Swinburne University of Technology	Victoria, Australia
University of Tasmania	Tasmania, Australia
University of Toronto*	Ontario, Canada
Very Long Baseline Array (NRAO)	New Mexico, USA
Yamaguchi University*	Yamaguchi, Japan
Yebes Observatory (IGN)	Guadalajara, Spain

Entries with * provided multiple responses. Omitted respondents requested to remain anonymous.

Survey questions did allow for response from sites which had yet to adopt the VDIF format—and would then instead request comment on what has prevented their adoption—but this was not the focus of the study and thus did not yield significant data, as expected.

A typical survey experience, for an institution that had adopted the VDIF format, would begin by asking which formats were currently in use at their institution: VDIF Legacy (16B headers), VDIF (32B headers), a public VDIF variant such as CODIF[2], and/or a proprietary or internal VDIF variant. The participant would then be asked what year they first adopted a VDIF-based format and for details about the hardware and software used in their institution to produce and/or analyse VDIF data. They would then be asked whether they were aware of any issues encountered within their institution relating to either a) their own production or analysis of VDIF data, or b) incorporation or analysis of VDIF data from another institution. Where issues were known, details were requested about both the target issues and any fixes or workarounds employed to address them.

The survey design then diverged from structured questioning, to request any details or anecdotes respondents were willing to provide that were relevant to the use or adoption of the VDIF format, as well as samples of VDIF-formatted data their institution has produced. Some participants—and even non-participants who did not view themselves as being part of the target demographic for the survey—also elected to send detailed reports on their experiences working with VDIF data to the organiser of the study via email.

Analysis of Responses and Sample Data

Survey responses included a mix of quantitative information—e.g. VDIF formats used, year of adoption—and quantitative data in the form of unstructured text responses. Thematic analysis of long-form responses was performed in a deductive fashion [7], beginning with themes identified from preliminary findings in sample data analysis and informal consultation with expert VDIF users. Though the design and motivation of the survey was initially based

[2] CODIF is the CSIRO Oversampled Data Interchange Format, a VDIF derivative that uses 64-bit words, 64-byte headers, and an extended set of metadata header fields to represent data incompatible with VDIF—such as from high sample rates among newer large sensor arrays. At this time, the CODIF format remains unratified, and its specification is not yet public.

on anecdotal evidence from the researcher's own work—issues she had analysing data from her own and other institutions, and challenges reported by her colleagues—analysis of responses did allow for expansion beyond anticipated themes, and this was ultimately found to be the case.

Identification of un-reported format compliance issues in provided VDIF samples was done with a Python-based tool which was developed as part of this study. At the conclusion of the study, this will be made available at https://github.com/TheMartianLife/VDIFVerify for public use and adaptation.

Prominence, Severity and Impact

Format Adoption

Beginning with format use and adoption, survey responses showed that use of the deprecated VDIF Legacy format is still widespread, and that the majority of those to adopt a VDIF format only did so 5-10+ years after its ratification. These numbers reinforce comments later made by respondents that technology in the domain of VLBI is often difficult and slow to change—whether due to highly specific hardware requirements or lack of developer time. Three institutions also reported using independent internal VDIF variant formats; though they did not provide specific motivations for the development of these variants, this does work against the stated goals of VDIF to be universal and consistent. Similarly, others reported that VDIF alone did not support features required by their institution, and that their use of the format relied on external metadata files.

Technical Context

The hardware that respondents used to produce VDIF data was relatively homogenous—settling on systems with ROACH Digital Backend (RDBE) [8] or Digital Baseband Converter (DBBC) digitisers [9] and Mark5 [3] or Flexbuff data recorders[3]—but the software each used to analyse or decode VDIF data varied widely. Responses raised concerns for the consistency and accuracy of decoded VDIF data, including:

- Several users reported having developed their own internal tools for VDIF processing. Therefore, the validity and correctness of these tools cannot be verified.
- Others reported using Open Source tools that vary in small ways in their interpretation of identical VDIF data, such as popular libraries mark5access[4] and baseband using different key values for decode operations[5].
- One user, in private correspondence, also noted philosophical differences in how they and a colleague each decoded identical data—with or without implicit 0 values for each n-bit encoding.

It should be noted that because this data was captured by some device and then encoded into the VDIF form that is being processed, there is only a single *correct* decoding—in as far as it accurately recreates the original information stream.

---

[3] Developed as part of the European VLBI Network's Novel Explorations Pushing Robust e-VLBI Services (NESPReS) project. See https://doi.org/10.22323/1.162.0033

[4] See https://svn.atnf.csiro.au/difx/libraries/mark5access/trunk/

[5] See https://github.com/mhvk/baseband/blob/.../baseband/base/encoding.py#L12

Format Non-Compliance

Concerns around internal and unverified tools only increase as the survey moved into questions around compliance issues. Almost a third of respondents self-reported ways in which their data did not conform to the VDIF spec, and even more had encountered non-compliance in files from other stations. Reports of issues spanned:

- **Timing issues**—incorrect reference epochs leading to incorrect handling of leap seconds, in-second frame numbers not resetting, incorrect time representations, more.
- **Encoding issues**—reversed byte ordering, inverse sign on complex data imaginary component, incorrectly listed number of channels, more.
- **Missing information**—omitted station codes, zeroed timestamps, number of channels or bits per sample always set to 1, more.
- **Pollutants**—each frame containing junk values towards the end to fill to some pre-set length, every frame has invalid flag incorrectly set, more.

In almost every case, the stated cause was either a recorder firmware issue or unknown, and the solution was some variant on manual modification of the output. Responses spoke widely of internally developed Python tools, private forks of popular libraries, or aftermarket firmware changes to backend components. As was noted about data processing software, the validity and correctness of these internal tools cannot be verified—and may, in fact, be introducing further inconsistency or error.

Several minor issues were identified in the provided data samples, beyond that which were reported. This is significant as only small portions of data were requested, so categories of issues that progress over time cannot be identified—such as issues which occur at second boundaries only. Respondents also commonly reported having provided specific data files that represented some notable experiment, supporting the idea that each likely provided data which had already been analysed elsewhere—and thus any critical issues may have already been addressed. The unreported issues followed similar trends to those noted in the responses, primarily missing or incorrect metadata values.

Users also provided several reports of past issues which had now been resolved, such as with hardware upgrades or upstream changes. There was no explicit mention of these issues being recorded in any way that future consumers of the affected data could reference.

User Attitudes

Aside from the issues and technical information reported by respondents, there was a notable disparity in the level of detail and apparent confidence exhibited in the survey responses. Individuals known to the organiser of the study as developers of popular VDIF-related tools spoke with clarity and depth about issues with the format and their use of it, while others made comments that betray a very different level of understanding. Several respondents noted variants on three main points of view:

1. **They did not know what their VDIF data looked like.** If there were issues, they would not know unless something was highly evident in the decoded and post-processed output or ultimate scientific findings.

2. **They did not know what *correct* VDIF data looked like.** When it came to identifying issues in their own data, it was very outcome-focused—changing whatever was needed to make their data "look right" or work with standard tools.
3. **They did not know where to get additional information.** Anything not provided in the specification document was unknown to them; they unknowingly duplicated existing tools, or filled in gaps in their knowledge with assumptions.

Similarly, the reporting of issues with the VDIF format was interlaced with reporting of issues that arise when processing VDIF with specific tools. This may demonstrate a lack of understanding of the format and its purpose among respondents.

Conclusion

The VDIF specification describes a strict formatting standard that exists to serve a complex and highly technical field. But findings from this study show that its adoption has not been without challenges. Some of these challenges are technical—such as limitations or differences in hardware or firmware—but many are organisational, including lack of expertise, time, public information and resources to effectively make use of the format and develop the necessary supporting tools for integration into existing workflows.

Issues reported by respondents in this study were rarely in conflict with the specification of the VDIF format itself; in fact, the majority spoke very highly of the format—both in isolation and in comparison to the solution they had used prior. However, individuals consistently reported issues with the non-technical aspects of standards adoption: how this format and other industry standards like it were communicated, integrated, evolved, and verified. Several respondents noted strong reliance on their professional networks to seek information and validate technical implementations that they required in their day-to-day operations.

Some issues present in the provided data and responses were sufficient to dramatically misrepresent the original collected data, though luckily there are no known cases where this has led to a false finding or significant impediment to a researcher that has used decoded VDIF data. It is more likely these issues would manifest by making additional work for data consumers to track down the necessary metadata or, in the worst case, producing junk data. Nonetheless, these issues should be documented alongside the affected data in archival, to ensure the required context is retained.

Future Work and Recommendations

Findings from this survey raised key questions for potential follow-up investigation, such as where and how users of technical standards such as VDIF seek information to stay up to date with evolving formats and issues, and where in their institution the responsibility for that lies. Further investigation of compliance issues in the VLBI data domain would ideally leverage mixed methods beyond survey alone, to better identify the necessary contextual information for individual solutions to be prescribed—such as the organisational drivers of success and failure to comply with specifications across various institutions.

The verification tool created during this study works to empower VDIF users to identify non-compliance issues in their own institutions and those with which they share data. But it cannot be assumed that lack of awareness is the only barrier to compliance; once identified, issues may remain unresolved without other forms of user support or governance. Adoption of

practices from other domains may be key to ensuring data consistency and compliance in the long-term. For example, standard practices from the Open Source software domain [10] that allow users to:

- **Affect change**—such as through community-led format development, in the form of evolution or extension proposals.
- **Request support**—such as through community leaders and integration resources, for potential users or those new to the domain to become familiar with the necessary tools and conventions.
- **Seek authority**—such as through a canonical source of ownership and authorship for the format, and avenues for contacting those individuals or entities for any required clarification.

Users may also benefit from a public record of past issues encountered by others, and the solutions they employed. Though some of these needs may be currently satisfied by associated entities—including the Global VLBI Alliance or the International VLBI Service for Geodesy and Astrometry—there is no such VDIF-focused entity that is both enduring and publicly-accessible. Given suitable representation from the diversity of VLBI institutions and instruments in operation, such an entity would be perfectly positioned to monitor and ensure compatibility between the data and practices of operators around the world. Until such an authority exists, incompatibilities between data producers are only being discovered by chance or individual interrogation.

Particularly as we look to the future—and begin to formalise designs for the next generation of VLBI data formats such as CODIF—it is essential to understand the strengths and weaknesses of existing data formats such as VDIF as they are used around the world today, by users with diverse expertise, infrastructure, and requirements.

Acknowledgements

The author wishes to acknowledge the contributions of each of the respondents to the survey described in this paper, as well as those who have worked to develop and administer the VDIF format specification. Many who took part well exceeded expectations in the level of detail they provided about their systems and work patterns, and the author was gladdened by the significant number of participants who noted that helping others in the field was a primary motivator for their taking part.

References

1. Harald Schuh and Dirk Behrend. 2012. VLBI: A Fascinating Technique for Geodesy and Astrometry. *Journal of Geodynamics* (61), 68–80. doi: 10.1016/j.jog.2012.07.007

2. Tatiana Bocanegra Bahamon, Leonid Gurvits, Guifré Molera Calvés, Giuseppe Cimò, Dmitry Duev, and Sergei Pogrebenko. 2019. VLBI and Doppler Tracking of Spacecraft for Planetary Atmospheric Studies. In *Proceedings of 14th European VLBI Network Symposium & Users Meeting (EVN)*. doi: 10.22323/1.344.0060

3. Alan Whitney. 2003. Mark 5 Disk-based Gbps VLBI Data System. In *New technologies in VLBI* (306), 123–134. Source: http://adsabs.harvard.edu/abs/2003ASPC..306..123W

4. Mark Kettenis, Chris Phillips, Mamoru Sekido, and Alan Whitney. 2009. VLBI Data Interchange Format (VDIF) Specification. Technical Report. Source: https://vlbi.org/vlbi-standards/vdif/

5. Pablo de Vicente. 2017. Technical Status and Developments of the EVN. In *Proceedings of the 13th European VLBI Network Symposium & Users Meeting (EVN)*. Source: https://iaaras.ru/en/library/paper/1679/

6. Alan Whitney, Mark Kettenis, Chris Phillips, and Mamoru Sekido. 2010. VLBI Data Interchange Format - An Overview. In *Proceedings of the 6th General Meeting of the International VLBI Service for Geodesy and Astrometry (IVS)*. Source: https://ntrs.nasa.gov/citations/20110011868

7. Benjamin Crabtree and William Miller. 2022. Chapter 12 – Template Organizing Style of Analysis. In *Doing qualitative research*. Sage Publications. ISBN: 9781506302812

8. A Neill, M Bark, Christopher Beaudoin, Walter Brisken, H Ben Frej, S Doeleman, S Durand, M Guerra, Alan Hinton, M Luce, R McWhirter, K Morris, G Peck, M Revnell, A Rogers, J Romney, Chester Ruszczyk, Michael Taveniku, R Walker, and Alan Whitney.2010. RDBE Development and Progress. In *Proceedings of the 6th General Meeting of the International VLBI Service for Geodesy and Astrometry (IVS)*. Source: https://ntrs.nasa.gov/citations/20110011812

9. Gino Tuccari, Salvatore Buttacio, Gaetano Nicotra, Ying Xiang, and Michael Wunderlich. 2006. DBBC – A Flexible Platform for VLBI Data Processing. In *Proceedings of the 4th General Meeting of the International VLBI Service for Geodesy and Astrometry (IVS)*. Source: https://ivscc.gsfc.nasa.gov/publications/gm2006/

10. VM Brasseur. 2018. Chapter 9 – When It Goes Wrong. In *Forge Your Future with Open Source*. Pragmatic Bookshelf. ISBN: 9781680506389