Gap Analysis Report on Incoherent Scatter Data

COOPEUS Deliverable 2.2

Prepared by Anders Tjulin[a], Phil Erickson[b] and Ingemar Häggström[a]
([a]EISCAT Scientific Association, [b]MIT Haystack Observatory)

2013-05-20

1 Introduction

The high-latitude atmosphere and ionosphere are important for studies of the relationship between Solar and Terrestrial conditions as well as the coupling of the different altitude regions in the Earth's atmosphere. That is why a large amount of effort has been dedicated to these studies in the polar regions. Since the systems that are studied are global in their nature, and not limited to national borders or political borders, the full view of the atmospheric and geospace environment is only possible through international collaborations.

Incoherent scatter radar (ISR) systems are important research tools in the studies of the upper atmosphere and the ionosphere, and the standard high-level data from these systems contain electron density, electron and ion temperatures, and line-of-sight plasma flow as functions of time and altitude.

There are about a dozen active ISR systems in the world at the present. Three of these systems are operated by EISCAT Scientific Association[1] and they are located in the northern Fenno-Scandinavia and on Svalbard with transmitters/receivers in Tromsø (Norway) and Longyearbyen (Svalbard) and additional receivers in Kiruna (Sweden) and Sodankylä (Finland). A next-generation ISR system, called EISCAT_3D[2], is now in the Preparatory Phase and is planned to be going into commissioning within six years from now. The corresponding US atmospheric ISR systems are Millstone Hill[3] in Massachusetts (run by MIT), Jicamarca Radio Observatory[4] in Peru (run by Cornell), and Arecibo[5], Sondrestrom[6] and AMISR[7] (the Advanced Modular Incoherent Scatter Radar, run by SRI).

The European and American ISR communities are relatively well integrated already with, for instance, incoherent scatter World Days, coordinated through the International Union of Radio Science (URSI), which is a programme of coordinated observations in order to provide a data set of synoptic ionospheric parameters on a global scale. There are about 20 World Days per year, scattered through the seasons. There are also common workshops and meetings, and the

[1] www.eiscat.se
[2] www.eiscat3d.se
[3] www.haystack.mit.edu/obs/mhr
[4] jicamarca.ece.cornell.edu
[5] www.naic.edu
[6] isr.sri.com
[7] amisr.com

Madrigal database system is used as standard repository for high-level ISR data from the different systems around the world. However, one of the aims of Work Package 2 of the COOPEUS project is to integrate the different worldwide ISR systems more in terms of standards and procedures for lower level data storage and data sharing.

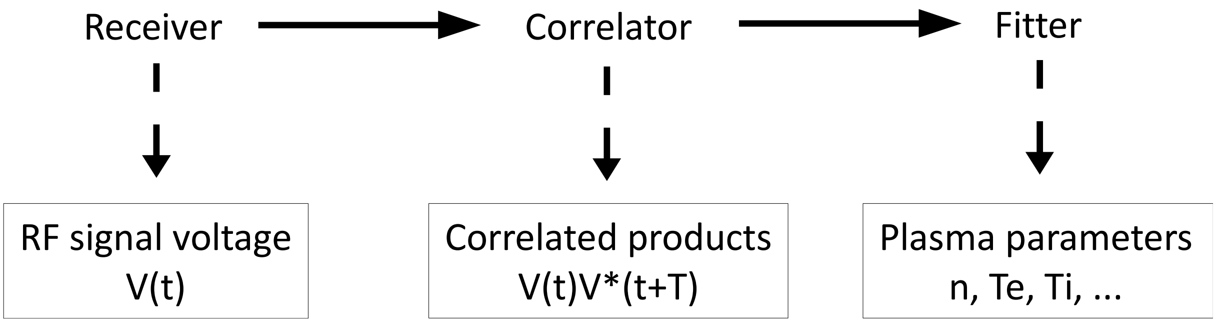
The goal of the present gap analysis has been to analyse the data practices at different ISR systems in the project in order to identify areas where harmonisation activities are needed. This document is a result from initial discussions between EISCAT (Ingemar Häggström, Ingrid Mann, Anders Tjulin) and MIT Haystack observatory (Phil Erickson, Frank Lind).

2 Different levels of data

ISR systems produce vast amounts of raw data which can be stored and transferred at different levels of reduction. There are many similarities between the different ISR systems in the world, and this opens up possibilities to standardise the data level formats for simplified scientific collaboration between the systems.

2.1 The signal chain

The typical flow for the signal from ISR observations can be described in block form as follows:



For simplicity, the diagram only shows the interface boundaries that are generating data products typically used for long term storage. The individual processing elements often have other intermediate products as a result of their internal calculation chains, but these are usually transient in memory or across a network and not useful for storage.

The software signal patterns for ISRs has been discussed by Grydeland et al. (2005) on a general level, and they are also discussed in Section 5 of the Open Radar Initiative RF Signal Format Standard Definition document[8]. The Open Radar Initiative is a project initiated in order to develop technology for radio science applications that is both reliable and reusable.

For the purposes of this discussion, the typical signal chain shown above uses the following two major software pattern classes during its operation:

Transformation: Elements that extract information from their input and pass that information along to output, usually through signal processing operations. These elements often provide a significant amount of data reduction with the trade-off of information loss from the original

[8] www.openradar.org/data/attach/Documents(2f)OpenRadarRFSignalObject/attachments/rf_signal_object.pdf

signal. Naturally, that information loss is arranged in a manner so that it does not remove information needed for later stages in the signal processing chain.

Weaving: Elements that increase information content by, in addition to their inputs, adding time or space aligned metadata. This additional metadata is applied in order to enable later stages in the signal processing flow to correctly interpret the contained information in the signal.

There are three typically persisted data formats in the ISR signal processing, corresponding to the different blocks in the illustration above. These are:

Radio frequency signal data: This is a *voltage level* representation of a radio frequency (RF) signal, as captured by a single digital receiver channel. Sufficient metadata associated with the signal, source, processing, and coherence are incorporated at this level. These metadata are intended to be sufficient to understand the context of the signal from the data itself.

Correlation product data: This is a *correlation level* representation of RF signal data after pairwise auto- and/or cross-correlation has been applied, in space and/or time. Most usually, some amount of temporal averaging has occurred at this stage. Some systems also, at this stage, perform spatial averaging such as application of summation rules to full-resolution lag profiles. Significant amounts of metadata are also added in order to characterise system parameters such as the beam position and calibration constants for instance.

Fitted plasma parameter data: This is a *plasma parameter level* representation of the ionospheric state as derived from correlation level products. This stage is intended to be the final measurement value, with estimates of the uncertainty, for the parameters analysed as a function of space and time. In certain signal chains, fitted parameter records representing basic or fundamental parameters, such as electron and ion temperatures and line-of-sight velocity, may be combined across time delimited records to produce derived parameters, such as neutral temperature and vector velocity, as another fitted plasma parameter record output.

The international ISR community is beginning to converge on uniform standards and descriptions for these three persisted formats, but the level of the efforts vary widely at each of these three levels within the major ISR systems in use today. In the following sections we discuss the current status of these harmonisation efforts at each of these three levels of data.

2.2 Radio frequency signal data

The radio frequency signal data format is well-defined at each ISR facility since it is simply the raw voltage level data from the receivers. However, this also means that the format is system dependent to a high degree. Until relatively recently, RF signal voltage data space requirements were sufficiently burdensome that systems often did not persist this data, which meant that there was no need for harmonisation of the data at this level. However, with advances in digital receiver, computing, and storage capability, many active ISR systems are moving to routine storage of RF signal data and therefore a need for a standardised format at the RF signal level has arisen.

EISCAT

At the EISCAT ISR systems it is possible to save the voltage sample streams. However, this is not done routinely at the present because of the large storage requirements. The experiment files are also needed here in order to understand from where the samples in the data come. The raw data are saved only for the cases where they would add some more insight to the data such as when running the VHF radar in interferometric mode. This data format from the EISCAT systems are only usable for a handful of very expert users of the system. If an EISCAT user wishes to have raw data from an experiment, there is a possibility to connect their own data recorders to the radar system at the stage from where they want their raw data. This process should certainly be more streamlined at the time when the EISCAT_3D system starts operating.

Millstone Hill

The Millstone Hill ISR has retained all RF signal voltage data for experiments since December 2001. All data processing in the Millstone Hill system starts with these data as the fundamental input. The stored bandwidths are typically about 100 kHz around the centre frequency which is sufficient for most ion line measurements. The data are stored in baseband I/Q format in a continuous manner with fixed size records and a minimal data header. At present an internal data format is used for the production MIDAS-W software radar system, but a robust and well defined HDF5 export file format[9] is available. Within the HDF5 file, separate branches contain baseband I/Q data (the RF signal) and metadata specific to Millstone Hill with information on the system state such as the selected antenna, transmitter power, and timing mode.

The fields and formats for these items are documented separately in the Open Radar Initiative RF Signal Format[10] and Open Radar Initiative Millstone Radar System Status[11] documents already provided to the COOPEUS effort. External feedback on the Open Radar HDF5 format was collected from several end users of RF signal data during its development, and was incorporated into the draft standard. The end users involved have included those whose primary experience is with hard target radar measurements.

AMISR

The AMISR systems have the capability of generating and storing RF signal voltage data using a dedicated RADAC card in a proper mode. The bandwidths of the stored data vary. An HDF5 export file storage format has been used for some time, and is well tested with external users of direct RF signal information. The Millstone Hill HDF5 standard was in fact based on initial experience with AMISR HDF5 formats, and was designed to provide largely the same user experience.

[9] www.hdfgroup.org/HDF5/

[10] www.openradar.org/data/attach/Documents(2f)OpenRadarRFSignalObject/attachments/rf_signal_object.pdf

[11] www.openradar.org/data/attach/Documents(2f)OpenRadarRFSignalObject/attachments/millstone_system_status_object.pdf

2.3 Correlated product data

Correlated products (or the frequency domain equivalents of auto- and cross-power spectra) are produced by signal chain elements applying auto- and cross-correlation operators to RF signal voltages. Typically, some amount of time averaging of these correlation products occurs as well. An estimate of the variance on the correlated products is also produced at this stage, as it is necessary for proper data weighting and fitting later on. A specification of the radar ambiguity function is included here, since it has information about the instrumental distortion introduced by the transmitted waveform and the RF signal chain's effective impulse response.

The data volume expands significantly for the returns from each radar sweep, defined here as a transmitted pulse plus the ionospheric return, due to the need to form many correlation lag products in the time domain (or frequency bins in the frequency domain). However, the data are reduced in volume when successive radar sweep lag products are time averaged, most usually through mean or L1 (e.g. median) operations. If M lags are computed but N radar sweeps are time averaged, then the data expansion factor is M/N , and this can sometimes be much less than unity.

In some systems, further data processing is applied to individual time averaged lag products.

Summation rules: A transform averaging together those lag products that are considered to be from an ionospheric volume possessing the same plasma parameters (i.e. one which is statistically stationary in a spatial and temporal sense, where the temporal stationary property has already been invoked using time averaging).

Lag profile inversion: An inverse approach which attempts to remove the effects of transmission waveforms and receiver impulse responses, to produce a lag profile which can be analysed assuming essentially an ideal radar ambiguity function. This technique is also known as lag profile regularisation in the literature.

Regardless of which operation is chosen, the output is still a lag profile matrix which means that the data container format is similar in most cases, but with extra metadata added containing information on the additional signal processing effects.

Persisted correlated product information has been used in active ISR systems for far longer than RF signal voltage persistence, but no uniform format for these correlated products exists.

EISCAT

The EISCAT ISR systems use GUISDAP[12](Grand Unified Incoherent Scatter Design and Analysis Package) as the container for correlated product storage. The GUISDAP experiment specification was introduced in the mid 1990s (Lehtinen and Huuskonen, 1996). Nowadays most EISCAT data, even from the early days, are described in GUISDAP, and it is the standard format used for long term storage of the data.

GUISDAP is a program package that is at the present written in MATLAB 6, and it includes the description of the data files. It is a binary format compatible with the “.mat”-standard file

[12] www.eiscat.se/groups/Documentation/UserGuides/GUISDAP/gup87.html

format as it was defined in MATLAB 4. The data files are intended to be directly read by the MATLAB software included in the GUISDAP package during the analysis.

The metadata corresponding to the correlated data are saved in a parameter block. The format of these blocks have changed somewhat over time. The amount of metadata is just sufficient for the analysis but should ideally contain more details. For instance, at the present only one value of the transmitter power and the antenna pointing direction is recorded per data dump, which normally takes on the order of five seconds. This time resolution is not sufficient when the power level fluctuates (which happens if the EISCAT Heating facility is used in connection to the measurements) or when the antenna direction is changed continuously during the measurements.

Millstone Hill

There are two formats in use at the Millstone ISR system.

The *CMST* format (Command/Status) container has been used for correlated product storage since the MIDAS-1 system came online in 1992–1993, and it is still the persistence format for the current MIDAS-W and MIDAS-M software radar systems. This data format was developed jointly with EISCAT at that time. The CMST representation is an XDR endian-agnostic binary format to support the transport of information across networks. Its implementation is complex, and Millstone Hill discussions have centred on an upgrade of this format to an HDF5 base or similar in the future.

The *IFAM* format is also in use. After summation rules have been applied, a single IFAM record is produced either in memory or (rarely) persisted to disk, as the container for input to the main INSCAL incoherent scatter fitting program. IFAM records are intended to contain all relevant metadata and auto-correlation functions needed to accomplish height-by-height determination of plasma parameters, including system state metadata and radar ambiguity function representation. However, this format is not intended for more complex processing such as full profile analysis where a CMST input is more suitable. IFAM specification exists only in Fortran and Python code at the moment and not as a formal document.

2.4 Fitted plasma parameter data

The ISR community has a long established data container for storage of fitted plasma parameters as a function of space and time. The Madrigal distributed database system[13], whose development started in 1980 at Millstone Hill, is now the standard community repository for all ISR data systems. The source code is distributed in an open source manner through the OpenMadrigal initiative[14].

In the Madrigal system, data are divided into different experiments, and each experiment has a defined overall start and end time. Catalogue and header records provide overall metadata such as experiment and site descriptions. Within an experiment, data are organised into individual records, each with its own start and end time. A record can contain one-dimesional values (single per record) and two-dimensional values (vector of parameters). The

[13] www.eiscat.se/madrigal

[14] www.openmadrigal.org

specification of the physical units for each value is provided as well, and most parameters also include an associated uncertainty where appropriate.

A derived parameter engine inside Madrigal allows additional parameters to be provided to users alongside the fitted parameters where appropriate (e.g. neutral temperature), and the engine also provides metadata values (e.g. background magnetic field, IRI reference ionospheric model parameters) as an aid for science analysis and interpretation. The derived values are limited to expressions that can be calculated within a single experiment record (i.e. time-limited), potentially using available external parameters within Madrigal itself. Derived parameters requiring more complex calculations are calculated by separate programs, which then create their own new Madrigal experiment files. An example of these parameters might be data products that synthesise information from multiple measurement records.

The underlying record data format was defined by the US CEDAR community in the late 1980s and is based on 16 bit integers, with significant restrictions on precision. The Madrigal system is moving within the next year or two to an underlying HDF5 flexible format in order to remove these limitations of the format.

While it is true that all ISR sites produce fitted data in Madrigal form, the structure of each site's data record differs widely. For example, Millstone Hill provides not only one-dimensional uncertainties in each plasma parameter, but also two-dimensional parameter correlations (i.e. correlation of T_i with T_e). These are not yet commonly used by the ISR community, but are a partial step toward providing the full covariance matrix for each measurement as a means of conveying the maximum amount of information about measurement “confusion”. The ISR sites may also use different parameter type codes for what is essentially the same measurement, although this situation has improved considerably in recent years as a result of careful Madrigal and CEDAR database curation.

The organisation of the records is also different within particular experiments, especially when considering different transmission waveforms producing different effective range resolutions in final data. For example, Millstone Hill at the moment produces a single batch calibrated file in which records cycle through all interleaved/simultaneous measurements at the same start/end time, while other systems such as AMISR produce separate experiment files for each waveform or equivalently each effective range resolution. Based on experience from teaching at ISR summer schools, the latter approach is often preferred by the users. A more uniform experiment organisation across the different ISR facilities would significantly ease the intellectual burden for a new user when trying to navigate the various measurement result files.

2.5 Discussion

The data most requested by the users are by far the fitted plasma parameters, that is the derived physical data. Less than 10% of the users request raw voltage level data, and very few are interested in data at intermediate level such as the correlated product data.

The raw voltage level data is the most expensive data to store and transfer, but it is useful in specific experiments and for developing new measurement and data analysis techniques, so it would be useful to standardise the format for easier exchange between users and systems. The

HDF5 experience of AMISR and Millstone Hill provides a place to start for discussions aimed at defining a common RF voltage format across the community.

The data at intermediate levels are the most convenient form of data for exchange and storage, because the data volume is decreased to a more manageable level while still retaining sufficient information for further analysis. By storing these data it is also possible to recalculate plasma parameters at later stages following possible new understanding of the upper atmospheric processes. Today there is no standard data format at this level so this is the area that needs most work for the harmonisation between the different ISR systems. We have to remember that this format works well for internal use at all present systems, it is when interacting with other systems that problems may arise at present.

Following an examination of the specifications at Millstone Hill and EISCAT, common features of these correlated product formats have been identified. These could be used as a starting point for a common format for correlated data. This is an initial list and would need to be expanded and corrected in a discussion process within the ISR community. Experience has also shown that it is useful to have a record of how system parameters were commanded separately from how the system actually performed; this implies though that actual system state needs to be recorded during experiment operations for all key variables.

The list of common features for the correlated data at EISCAT and Millstone Hill follows:

Common features of correlated ISR data formats					
System metadata	Antenna – specified for TX and RX systems separately:	Site reference coordinates			
		Number of beams			
		For each beam:	Beam shape		
			Beam pointing		
			Beam pointing coordinate system		
			For each feed polarisation:	Feed bandpass shape	
				Feed amplitude and phase calibration	
	Transmitter:	TX frequency time history			
		Modulation envelope time history:	Amplitude		
			Phase		
			Peak power		
			Average power		
	Receiver:	Number of channels			
		For each channel:	Association of that channel with antenna beam / feed / polarisation, or with sampled TX channel		
			Receiver centre RF frequency		
			Receiver final frequency (=0 for baseband data)		
			Receiver effective impulse response / bandpass shape (amplitude and phase)		
			Receiver timing:	RF blank interval, if present	
				Attenuation time history	
				Noise calibration diode injection system parameters (if present):	Timing of noise calibration pulse injection
					Absolute noise calibration pulse power
				Receiver amplitude and phase calibration	
				Receiver noise temperature	
Correlated product data	Number of lag profiles				
	For each lag profile:	Associated receiver channels (different if cross-correlation, same if auto-correlation)			
		Associated transmitted modulation			
		Radar ambiguity function: delay × lag × range			
		Time average interval			
		Start offset from leading edge of TX pulse			
		Range sampling vector			
		Lag sampling vector			
		Lag product matrix (range × lag)			
		Lag product variance matrix (range × lag)			
		DC offset as a function of range			

3 Future plans within the COOPEUS project

The plans for future interactions between EISCAT and its US counterparts are to continue the video-conferencing. In addition, a two day workshop is planned to take place preceding the COOPEUS annual meeting in September.

4 References

Grydeland, T., Lind, F. D., Erickson, P. J., and Holt, J. M., “Software Radar signal processing”, Annales Geophysicae, **23**, 109–121, doi:10.5194/angeo-23-109-2005, 2005.

Lehtinen, M. S., and Huuskonen, A., “General incoherent scatter analysis and GUISDAP”, Journal of Atmospheric and Terrestrial Physics, **58**, 435–452, doi:10.1016/0021-9169(95)00047-X, 1996.