WEC Requirements Specification

The Joining of Cluster WEC Data

WECdata Joining Working Group

Tony Allen, Anders Lundgren, Michel Parrot,
Simon Walker, Chris Harvey (chair)

Contents

Document Status Sheet			
1. Document Title: **Joining of Data Sets**			
2. Document Reference Number: **CWD-OBSPM-TN-001**			
3. Issue	4. Revision	5. Date	6. Reason for Change
1	1	Sept 5, 1994	First issue
1	2	September 26, 1994	Change last paragraph of section 1 Add new paragraph at start of section 5 Correct indices in equation 3 and 4

1 Introduction

Joining of data streams is exceedingly important for Cluster; the primary objective of the mission is the comparison of data from four different satellites with independent spin phases and spacecraft clocks. Any two data sets must be "joined" for any comparison other than simple plotting on separate curves.

The distinction between "merging" and "joining" is defined in the document "CDF File Design for Cluster: Recommendations to CSDS", DS-QMW-TN-0001, dated 94/05/20. Basically, merging retains both time lines, whilst joining places all the variables onto a single time line. Merging and joining are quivalent for two data sets already on the same time line.

Closely related to the joining of data sets is their concatenation. Whereas joining concerns data from different sources acquired during the same time interval, concatenation concerns data from the same source for different, but contiguous, time intervals.

The purpose of this document is to describe the issues involved and outline possible methods of joining data sets. Requirements for WEC are proposed in section 9. These are minimum requirements which must be implemented. The use of more advanced techniques, such as Butterworth filters, has yet to be investigated. If they can be shown to be beneficial, there is no reason not to add later extra options to the menu of available joining algorithms.

2 The Use of Joining

JOINING of data is generally required when data comes from different sources. In particular, joining is required:

1. To compare high resolution data with the PPD and/or SPD, to compare the PPD with the SPD, and generally to compare any two data sets which are sampled at different frequencies.
2. To compare data from the different spacecraft, because their data acquisition is not synchronised.
3. To compare spin-sychronous and clock-synchronous data from the same spacecraft.
4. To compare satellite data with ground-based data and with data from other spacecraft.

CONCATENATION is required:

5. To produce files which are continuous over different segments of the source data files: typical examples are continuity over midnight, or over the end of a year.

3 Discussion

Any two data sets which are sampled at a different frequency will eventually become seriously out of step if step are not taken to avoid this. The joining of two data sets puts them both onto a common time-line.

For several types of analysis exact synchronisation of the sampling is required, but not exact simultaneity; but, in practice, any action taken to ensure synchronisation can easily provide simultaneity.

Synchronous sampling is essential for

- taking differences, *e.g.*, to calculate gradients between spacecraft, and
- most, but not all, numerical correlation of data.

It is not necessary for visual (i.e., graphical) correlation.

Concatenation is a relatively simple file copying routine, except:

- when one or more data points are missing, or
- when there is a discontinuity of the time line.

4 Joining Algorithms

Many different methods of joining can be imagined. The choice is determined by the requirements of the subsequent analysis. The two basic types of join are

1. Fuzzy join. Each datum of the data set with lower time resolution is associated with the datum closest in time from the data set of higher time resolution. The advantage of this type of join is its simplicity, while the main disadvantages are:
   - the quasi-periodic discontinuities which are introduced into the joined data set;
   - the possibility of undersampling of the data (see discussion in section 4.1.1).

   Fuzzy joining is not recommended for anything other than the simplest subsequent analysis.

2. Re-sampling. One of the data sets, which we call the secondary data set, is used to estimate the values of the same (*i.e.*, the secondary) variable at times which coincide with the times of the samples of the primary data set. This operation simulates resampling of the secondary data, which can be performed by:
   - averaging;
   - linear interpolation (see section 8 for examples);
   - higher order interpolation, such as spline functions.

   These are merely different methods of resampling.

   It may sometimes be desirable to use an artifical time-line instead of one defined by some primary data set; for example, samples once per second, on the second. This option may be attractive:

   - when several different data sets must be joined successively onto the same time-line;
   - to produce data sets spanning an arbitrary time interval but containing a pre-determined number of samples, *e.g.*, 2^n where n is an integer as reqired for some FFT programmes.
   - to concatenate files in the event of a timing discontinuity *e.g.*, if a spacecraft or an instrument clock is reset.

4.1 Precautions

Care must be taken when re-sampling.

4.1.1 Averaging

When decreasing the time resolution of the secondary data set, it is not adequate to interpolate between the two data points nearest in time to the primary set datum; the resulting secondary data set would be UNDERSAMPLED, and very dangerous to use. Rather, the resampling algorithm must include numerical filtering, so that all the data points are accorded their due weight in the computation of the re-sampled data set. Precisely how this filtering is performed depends on the subsequent use to be made of the data. In section 8 below two filtering algorithms are suggested:

- equal weight filtering, suitable for use when the resampled data is to be compared with data acquired using digital filtering techniques;
- weighted filtering, more suitable for correlation with data acquired using analogue filtering.

4.1.2 Interpolation

The primary data set will normally be the one with lower time resolution, so that joining reduces the time resolution of the secondary set. It is not generally useful to increase the time resolution of normal science data from an instrument which has been well designed with its pre-converter cut-off frequency close to the Nyquist frequency; physically meaningful information cannot be created by interpolation. There may be an occasional need to interpolate science data, when joining several data sets of high resolution and only a few of lower resolution.

- Linear interpolation is the only type of interpolation which may be applied to science data. The use of higher order interpolation on science data is dangerous: experimental errors and noise in the original secondary data set introduce completely spurious information into the resampled data. In any case, it is impossible to extract from normal science data any correlation with a precision in time greater than twice the length of the larger of the two sampling intervals of the original data (Nyquist's theorem).

  Data resulting from linear interpolation is OVERSAMPLED. It would be useful to append to the meta-data of the resampled data set an "oversampling factor" to indicate the factor by which the sampling frequency exceeds the requirements of the physical content of the signal. This flag would then be passed to application programmes and/or the scientists using them, to avoid fruitless searching for correlations where none is to be found.

- Higher order interpolation may be performed only on data which is oversampled (*i.e.*, data which is known to contain no spectral information at frequencies near the Nyquist frequency); for example, orbit data. Nevertheless, standard oversampled data products will normally be provided with sufficient time resolution for linear interpolation to be entirely adequate; this is true for the orbit/attitude SPD and PPD.

  If WEC has any requirement for higher order interpolation, an entirely separate interpolation application should be provided. A special flag in the meta-data to indicate the order of interpolation permitted could be useful to guard against unintentional or ill-advised misuse of this special interpolation routine; the default value would be 1.

5 Time Offsets

It has been suggested that ISDAT provide the possibility of inserting time offsets into a data set.

We must be clear what we mean by a time offset. The whole purpose of resampling of data is to produce samples at times different from those of the original data set; but when these new samples are plotted as a function of time, the curve obtained follows closely that of the original data set. The effect of introducing a "time offset" is quite different: the curve corresponding to the offset data would appear to be shifted forwards or backwards in time with respect to the original data set.

Although the introduction of time offsets would be easy to implement, this seems dangerous in the context of joining. All ISDAT servers and operators are aimed at enhancing the data quality. If there is a reduction in the information content, for example by averaging, it is reduced in a logical and understandable way, without reducing the data quality. But the introduction of a time offset is little short of the introduction of a deliberate error into a data set.

We may note that it is, in any case, difficult to define a precise time offset, for example, for data obtained on two different spacecraft. To determine a time offset, one normally thinks of cross-correlation: but in what band of frequencies ? The "offset" may well be frequency dependent, on account of propagation effects. The better way to determine any such "offset" if it exists, is to calculate the cross-spectrum of the correctly dated data streams; if a unique time offset can be defined, it will be seen from the relative phase differences increasing linearly with frequency.

Joining can be used to create data sets which are stored for later use. If time offsets are desired, they should be implemented at the level of the application, not at the level of the joining algorithm.

6 Meta-data and Status data

Meta data is information which is intimately attached to the physical data (units, coordinate system, possibly limit values). It must be joined when the physical data is joined.

Status data is information which is separated from the instrument data to which it refers, for use independently of the instrument physical data. In particular, it may be used by other experiments, to determine the conditions for their own observations. It has an inherent time granularity, which is somewhat arbitrary.

- Two files of status data may be joined. The length of the resulting status file is approximately equal to the sum of the lengths of the two input files (unless the two input files are synchronised in some way).
- Status data may be joined to science data (but the inverse has no sense); the status data becomes part of the science meta-data.

7 Implementation Issues

The implementation of an application to join data sets is intimately related to the structure of the data files used either for storage (*e.g.*, CDF), or analysis (*e.g.*, inside ISDAT).

Space physics data is characterised by being organised in long strings of data objects in chronological order. For each value of the time, the corresponding data object consists of one or more values which are all determined simultaneously. The data structuration presently being discussed within ISDAT separates:

- the time dimension, in which averaging and other more complicated processes like re-sampling or Fourier transformation can be performed,
- the space dimension, in which coordinate transformations may be applied, and
- other (non-space, non-time) dimensionality of the data set (such as energy or frequency).

The question of joining thus become one of forcing two data sets to use the same timing information. In ISDAT the time information is kept in the form of a table giving the start and stop times plus the number of records of the data file; in the case of time-tagged data, there is an additional array with the actual times of each data record.

If more than two data sets are to be joined, several secondary data sets can be resampled successively, one at a time, onto the timeline of the same primary data set.

When joining two data sets of similar time resolution, it clearly speeds conputation if the data set with the more complicated data structure is chosen as the primary data set. The end user must be made aware of this responsibility.

7.1 Open Issues

7.1.1 Where does the resampled output go ?

Should the resampled output stream be

- be physically merged with the primary data set, to create one stream on output; or should it
- continue its independent existence on its new timeline ?

If the resampled data set maintains its own separate existence, then the metadata must

- indicate the fact that it has been resampled, and
- identify the primary data set used to synchronize this resampling.

The ISDAT architecture consists of the data base handler which is a source of data, and filters which take take one stream of input and use it to produce one stream of output. If the data is physically merged, two data streams are used on input, and only one is produced on output. Joining is thus "topologically" quite distinct from other ISDAT operations.

7.1.2 Unused timelines

Data from the secondary data set is interpolated onto the time tags of the primary data set. The join is performed by applying, for each time tag of the primary data set, the same interpolation algorithm to each of the elements of the appropriate data objects of the secondary data set. After the joining has been performed, the timing information originally associated with the secondary data set is no longer needed if the secondary data set is no longer needed; can the associated time-line be discarded ?

This is an archtectural design issue (*e.g.*, there may be other data sets pointing to the same time-line).

7.1.3 Interface with ISDAT data base handler

Presently, ISDAT allows the user to specify the maximum number of points to return for a given data request. If the number of points to extract exceeds this limit, then the user specifies what action may be taken to reduce the number of points, *e.g.* none or averaging.

How is this capability interfaced with the requirements of the present document ?

8 Resampling Algorithms

We wish to joint the primary data set

$$\ldots \quad T_{j-1}, X_{j-1}, \quad T_j, X_j, \quad T_{j+1}, X_{j+1}, \quad \ldots$$

and the secondary data set

$$\ldots \quad t_{j-1}, x_{j-1}, \quad t_j, x_j, \quad t_{j+1}, x_{j+1}, \quad \ldots$$

by resampling the secondary data set so that the resampled data point $\ldots \quad T_j, x'_j \quad \ldots$ coincides in time with the primary data point $\ldots \quad T_j, X_j \quad \ldots$. We assume that the times T_j and t_j refer to the centres of their respective sampling intervals, and proceed as follows.

For every data point T_j, X_j of the primary set, we identify two data points of the secondary set, m and n, such that

$$t_m \leq (T_{j-1} + T_j)/2 < t_{m+1} \quad \text{and} \quad t_{n-1} \leq (T_j + T_{j+1})/2 < t_n \ . \tag{1}$$

Thus the interval $t_m \leq t \leq t_n$ of the secondary data set completely spans the sampling interval of the primary data set associated with the time-tag T_j. All the data points t_i, x_i in the interval $m \leq i \leq n$ (except possibly the first) will contribute to the resampled value x'_j. Let

$$N = n - m \ . \tag{2}$$

Note that eq. 1 ensures that $N \geq 1$, and that the total number of points from the secondary data set used for resampling is $N + 1 \geq 2$.

In the following section averaging and interpolation algorithms are proposed, using either

- uniform weighting to simulate true (*e.g.*, numerical) integration of the secondary data set, or
- linear weighting, to numerically simulate analogue sampling of the secondary data set in a computationally simple way while respecting causality.

8.1 Averaging

Averaging is best used when the time resolution of the secondary data set is reduced by a large factor. It cannot be used for interpolation !

8.1.1 Unweighted

The calculation of the unweighted mean is straightforward:

$$x'_j = \frac{1}{N+1} \sum_{i=0}^{N} x_{m+i} \tag{3}$$

8.1.2 Linearly weighted

The computation of the linearly weighted average is almost as simple:

$$x'_j = \frac{1}{N(N+1)} \sum_{i=0}^{N} i x_{m+i} \tag{4}$$

8.2 Linear Interpolation

Linear interpolation introduces less noise near the Nyquist frequency, especially when the sampling frequency of the secondary data set is being reduced only slightly ot not at all; and interpolation is essential if the sampling frequency is being increased.

We fit the straight line

$$x = a + bi \quad \text{with} \quad 0 \leq i \leq N \tag{5}$$

to the $N+1$ data values. Then, the resampled data value is

$$x'_j = a + bN\frac{T_j - t_m}{t_n - t_m}\,. \tag{6}$$

The coefficients a and b are determined by least squares, using either uniform weighting or linearly increasing weighting. It will be seen that

- If $N = 1$ uniform weighting must be used, which reduces in this case to linear interpolation between the nearest two data points.
- If $N = 2$ the linear weighting reduces to uniform weighting (*i.e.*, linear interpolation) using the last two (of the three) data points.
- If $N \geq 3$ then there is a real choice to be made between the two methods of weighting for the linear regression line.

8.2.1 Unweighted

For an unweighted linear interpolation, we compute the sum of the squares of the displacements,

$$S = \sum_{i=0}^{N}(a + bi - x_{m+i})^2 \tag{7}$$

and determine a and b so as to minimise S. Thus, we must solve

$$\begin{array}{rcccl} \frac{\partial S}{\partial a} & = & 2\sum_{i=0}^{N}(a + bi - x_{m+i}) & = & 0 \\ \frac{\partial S}{\partial b} & = & 2\sum_{i=0}^{N}(a + bi - x_{m+i})i & = & 0\,. \end{array} \tag{8}$$

These equations are

$$\begin{pmatrix} N+1 & \frac{N(N+1)}{2} \\ \frac{N(N+1)}{2} & \frac{N(N+1)(2N+1)}{6} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{N} x_{m+i} \\ \sum_{i=0}^{N} i x_{m+i} \end{pmatrix},$$

which may be solved, thus

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{N(N+1)(N+2)} \begin{pmatrix} 2N(2N+1) & -6N \\ -6N & 12 \end{pmatrix} \begin{pmatrix} \sum_{i=0}^{N} x_{m+i} \\ \sum_{i=0}^{N} i x_{m+i} \end{pmatrix}. \tag{9}$$

In the case of $N = 1$ this reduces to

$$a = x_m \qquad b = x_n - x_m$$

and we have linear interpolation between x_m and x_n.

8.2.2 Linearly weighted

We compute the sum of the squares of the displacements, weighted linearly according with time, as follows

$$S = \sum_{i=0}^{N} (a + bi - x_{m+i})^2 i \tag{10}$$

and determine a and b so as to minimise S. Note that the weight accorded to the first data point, x_m, is zero, so that effectively it does not enter into the determination. To minimise S, we must solve

$$\begin{array}{rcl} \frac{\partial S}{\partial a} & = & 2\sum_{i=1}^{N}(a + bi - x_{m+i})i \quad = \quad 0 \\ \frac{\partial S}{\partial b} & = & 2\sum_{i=1}^{N}(a + bi - x_{m+i})i^2 \quad = \quad 0\ . \end{array} \tag{11}$$

These equations are

$$\begin{pmatrix} \frac{N(N+1)}{2} & \frac{N(N+1)(2N+1)}{6} \\ \frac{N(N+1)(2N+1)}{6} & \left(\frac{N(N+1)}{2}\right)^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} i x_{m+i} \\ \sum_{i=1}^{N} i^2 x_{m+i} \end{pmatrix},$$

which may be solved, thus

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{1}{(N-1)N(N+1)(N+2)} \begin{pmatrix} 3N(N+1) & -(4N+2) \\ -(4N+2) & 6 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{N} i x_{m+i} \\ \sum_{i=1}^{N} i^2 x_{m+i} \end{pmatrix}. \tag{12}$$

In the case of $N = 2$ these equations give

$$a = 2x_{m+1} - x_n \quad \text{and} \quad b = x_n - x_{m+1}\ ,$$

so that

$$a + b = x_{m+1} \quad \text{and} \quad a + 2b = x_n$$

and the method is seen to reduce to linear interpolation between the last two data points, $x_{m+1} (= x_{n-1})$ and x_n.

9 WEC Requirements

The WEC requirements for joining are summarised below.

9.1 Joining

1. A software application must be provided for joining any two (within reason) data sets. Multiple joins will be effected by joining the secondary data sets one at a time onto the primary data set.

2. This application must provide the option of executing either the averaging or the interpolation algorithm described respectively in sections 4.1.1 and 8.2.

3. The default algorithm will be that of section 8.2.1, unless the secondary data set is having its sampling frequency reduced by a factor of more than 5 (value TBC), in which case the default algorithm will be that of section 8.1.1.

4. The application may also provide as an option the replacement of these two algorithms by those described respectively in sections 8.2.2 and 8.1.2. These algorithms are more suitable for resampling data priot to for comparison with analogue filtered data sets.

5. If the resulting resampled data continues its independent physical existence, then its associated metadata must indicate (section 7.1.1):
   - the fact that it has been resampled,
   - the resampling algorithm which has been used,
   - the primary data set used to synchronize the resampling, and
   - the oversampling factor (section 4.1.2).

6. If there is a required to perform a fuzzy join, it should be via a different application (because a fuzzy join produces a fuzzy result, see section 4).

7. An independent time-line generator must be provided (section 4).

8. All applications using two or more data inputs (*e.g.*, correlation) must test to verify that no further resampling is required.

9.2 Concatenation

1. An application must be provided to concatenate two consecutive data files of identical nature.

2. This concatenation application must:
   - verify the continuity of the timeline;
   - be capable of linear interpolation of one or at most two missing data objects.

9.3 Non Requirements

This section is included precisely because it may be controversial !
No requirements have been identified for:

1. Interpolation of higher order than linear (section 4.1.2).

2. The introduction of time offsets (section 5).

9.4 Open Issues

1. All the questions raised in sections 7.1.1 through 7.1.3 must be resolved.

2. Leap seconds are appended to Universal Time approximately once every two years (to take account of the slowing down of the Earth's rotation). How do the CSDS and the WEC data systems handle leap seconds ?