The Use of the Cyclic Redundancy Check Algorithm to Generate Unique Data Tags for Ocean Profile Data

Abstract

The Global Temperature and Salinity Profile Programme (GTSPP) has developed and tested a procedure to generate unique data tags for original ocean profile data by using the cyclic redundancy check (CRC) algorithm and successfully incorporated the CRC algorithm into its daily data processing stream.

The analyses of the CRC algorithm have been done by the US National Oceanographic Data Center, who hosts the long-term archive centre of the GTSPP. Up to mid-April 2012, the results of the analyses illustrated that 72,141 (94.5%) of 76,355 stations received from real-time data assembly centres matched with the delayed mode data stored in the GTSPP long-term archive. The CRC was able to correctly resolve 116 possible exact and/or near duplicates that could not be matched by conventional means.

The GTSPP has demonstrated the effectiveness of using CRC calculation as a unique tag for upper ocean temperature data. And the evidence here demonstrates the usefulness of that unique data tag. This Guide describes how the CRC algorithm can be used to generate unique data tags for ocean profile data.

I. Introduction

One of the difficult problems faced by ocean profile data providers is to match real-time and delayed mode, also known as non-real-time, data [2] from the same original observation. The problem stems from the need to deliver data quickly to users, with a trade off in reduced quality. Typically, the real-time version has reduced vertical and measurement resolution. There can also be uncertainties or inaccuracies in the real-time data both in positions and times, and they may contain un-calibrated profile data. The delayed mode data, arriving some time later usually have these errors corrected and so matching real-time data to delayed mode is not simply a matter of matching ship identifier, position and time. The conventional method of matching real-time and delayed-mode ocean profile data from the same original observation by positions within 5 km distance and separated in time by up to 15 minutes is not always satisfactory because of, for example, calibrations performed on the delayed mode profile

To resolve these problems, using a cyclic redundancy check (CRC) calculation was first suggested to be used in the Global Temperature and Salinity Profile Programme (GTSPP) (http://www.nodc.noaa.gov/GTSPP/index.html) and discussed at a GTSPP meeting in Hobart, Australia, 2002. Based upon the strategic framework discussed at the meeting, GTSPP developed and successfully incorporated the CRC algorithm into its daily data processing stream.

This Guide is organized in the following manner: the problem addressed by a unique data tag (UDT) is described in section II, followed by the technical approach and design philosophy covered in section III. The implementation strategy and preliminary results of the UDT are

described in sections IV and V, respectively. The conclusion of the CRC algorithm is given in section VI.

II. Defining the Problem

Most ocean profile management centres have developed software that considers detailed comparisons of individual station data when real-time and delayed mode positions are within 5 km distance and separated in time by up to 15 minutes (These are the criteria adopted by GTSPP but other centres may use different values). When these conditions are met, the detailed measurements of the profiles are compared. In an ideal case, the measurements of the real-time profile will exactly match those in the delayed mode profile to the resolution of the real-time profile. When this happens, the software reports an exact match. However, exact matches are not always made because of changes in data values as a result of calibrations performed on the delayed mode profile. Corrections can alter both the pressures or depths reported and the measurements at those levels. In this case, the software assesses the degree of similarity between the two stations and if it is high enough, alerts a technician who then must make a judgment whether or not the real-time was derived from the delayed mode or not. In most cases, where there is some uncertainty, the technician will err on the side of keeping both profiles.

Corrections are also sometimes required to the position or time. If these corrections shift the position or time outside of the 5 km, 15 minute window described above, no match will be found and the real-time and delayed mode profiles will appear to be from different originals even though they are not.

Over time, the number of stations coming in real-time and unmatched to delayed mode data will accumulate. Some of these will be because the delayed mode data has not been received, but some will be because a match was not possible even when the profiles did originate from the same observations. The archive thus has duplicates that bias statistics and mistakenly report available data. It is this problem of matching real-time to delayed mode versions of the same data that is being addressed with the unique tag.

III. The Cyclic Redundancy Check (CRC) Solution

Calculation of a CRC is standard procedure in many communications systems used today. Any system that checks the integrity of data packets and requests re-transmission if the test fails is based on a CRC calculation. The CRC algorithm is well documented [1]. It can be used in 4, 8, 16 and 32 bit forms. In this implementation we use the 32 bit version. This results in an 8-character tag which is the hexadecimal result of the CRC calculation. In simple terms, the algorithm is provided a string of characters of any length, and it produces a result that has one chance in 2^{32} (for the 32 bit version) of being the same as another, different input string. The solution exploits this to assign a unique identifier to the original data and allows a reliable comparison of real-time to delayed mode data. It is absolutely essential that the same input results in identical output.

IV. Implementation of the CRC Calculation

Four countries are involved in this project to evaluate the usefulness of the CRC value to provide a unique tag. Australia is involved because it is their idea and they already use it in checking for exact duplicates. They also contribute data from XBTs to the GTSPP both in real-time and delayed mode. Canada's Integrated Science Data Management (ISDM) agency is involved because it receives all real-time data distributed on the Global Telecommunication System (GTS), carries out QC and duplicates checking and then sends the real-time data to the US NODC for archiving for GTSPP. The US Shipboard Environmental (data) Acquisition System (SEAS) program (http://www.aoml.noaa.gov/phod/goos/seas/index.php) is involved as they are the single largest contributor of XBT data to the GTS and to the GTSPP. The US National Oceanographic Data Center (NODC) is involved as the archive centre that accepts real-time data coming through Canada and needs to match these to the delayed mode data arriving from varied sources.

To work, the CRC calculation must be attached to the original profile that eventually arrives in delayed mode, but also be derivable from or carried with the real-time data that are received quickly. Figure 1 illustrates the flow chart of the CRC approach for data from SEAS. Other centres use slightly different processes. Details of the implementation are provided below and portrayed in figure 1.

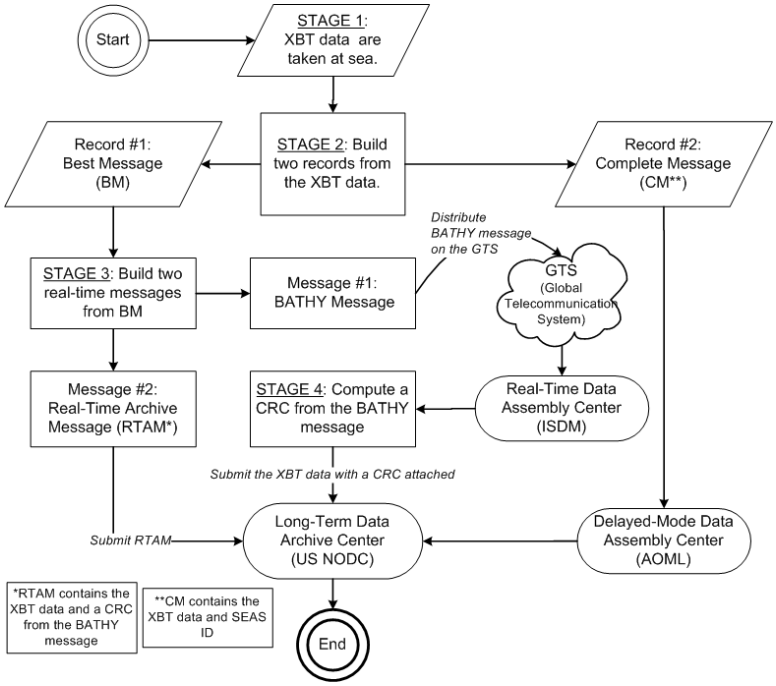

Figure 1 Schematic diagram of the Cyclical Redundancy Check (CRC)

STAGE 1: XBT data are taken at sea. After an XBT profile is collected it must be prepared for submission to both the Global Telecommunication System (GTS) (as a BATHY message) and to the archive centres, US NODC.

STAGE 2: Two records from the XBT data are created. The first one is called the “Complete Message” (CM) and the other is called the “Best Message” (BM).

In the SEAS implementation, CM is the XBT data stored on board the computer on the ship with a unique identifier created by the SEAS program with their own algorithm (call this the SEAS_ID). The SEAS software allows CM data to be sent ashore to a delayed-mode data assembly centre, Atlantic Oceanographic and Meteorological Laboratory (AOML) in full resolution and builds two real-time messages from BM in STAGE 3 as described below to be sent to the GTS and the GTSPP long-term archive centre (US NODC), respectively.

STAGE 3: Two real-time messages are generated from BM: (1) the BATHY message and (2) the "Real-Time Archive Message" (RTAM).

BATHY Message: The shore station constructs the BATHY message. The BATHY (or TESAC when CTDs are used to collect salinity data as well as temperature) message is sent to the Global Telecommunication System (GTS) and received by ISDM who are responsible for managing the real-time data for GTSPP. Because the BATHY code form has no place to transmit the CRC value, a real time data assembly centre has to be able to derive the CRC value from the BATHY message itself. The original version was coded by Australia and the code was given to other partners.

At ISDM, the same CRC calculation is made on the same portion of the message used to create the CRC value when the RTAM was constructed. The result will be identical to the CRC originally calculated as long as there has been no corruption of the message in GTS transmission. ISDM carries out QC on the real-time data, attaches the CRC they calculate and forwards these data to the US NODC. Note that in some cases, the position, time, ship identifier, sometimes even profile data, may be altered by ISDM as a result of quality control. In this case, the contents of the data received at ISDM may be altered, but the CRC tag is not.

Real-Time Archive Message: The XBT data also come ashore with the SEAS_ID. The shore station constructs the BATHY message and uses the contents of the BATHY (all characters including white spaces following the $8888k_1$ group of section 2 up to and including the = sign following the call sign at the end of the message; for a TESAC the CRC routine uses all characters after the $888k_1k_2$ up to and including the = sign) as input to the CRC routine. (For a TESAC the CRC routine uses all characters after the $888k_1k_2$ up to and including the = sign). An example can be found in Appendix A. The resulting 32-bit integer is converted to an 8 character hexadecimal number and this is attached to the original data as the CRC tag. The data that have come ashore, with the SEAS_ID and CRC tag are sent to the US NODC.

At this point, the US NODC has the real-time data from ISDM with quality control and the CRC tag. They also have received the XBT data that came ashore with no QC and with a CRC tag and the SEAS_ID. When the CRC tags match between the data received from ISDM and the data received from SEAS, they are considered exact matches. Later, the delayed mode version of the data arrives with a SEAS_ID only. NODC compares the SEAS_ID to that stored in the real-time archive to make a definitive match between the full resolution delayed mode XBT data and the real-time data sent over the GTS.

There are three rules that must be obeyed.

1. There is agreement on input to the CRC algorithm
2. The CRC tag is only generated one time and once generated it is never altered
3. All data exchanges include the CRC tag

V. Results

The analyses have been done by the US NODC since they are the agency where the real-time and delayed mode data meet. The match types are listed in Table 1.

Table 1 Description of matching real-time and delayed mode data

Match Types	Descriptions
A	Match by CRC and Exact match by Time and Position
B	Match by CRC and Inexact match by Time and Position
C	Match by CRC and No match by Time and Position
D	CRC present, but no CRC match; Exact match by Time and Position
E	CRC present, but no CRC match; Inexact match by Time and Position
Z	CRC present, but unmatched to data received from ISDM

From 25 February 2004 to 1 April 2012, NODC has received 76,760 stations from the SEAS program, with CRCs and SEAS_IDs. In examining the data they found the following results.

Match Type A: 65,785 stations (85.7%) received from SEAS and received from ISDM had identical CRC tags and were also identified as duplicates (exact match by time and position) using conventional software methods.

Match Type B: 6,356 stations (8.3%) received from SEAS and received from ISDM had identical CRC tags and were also identified as near duplicates by an inexact match of positions within 5 km distance and separated in time by up to 15 minutes.

Match Type C: 89 stations (0.1%) were found with matching CRC tags but with no match using conventional software techniques (stations within 15 minutes in time and 5 km in distance).

Match Type D: 103 stations (0.1%) were matched as exact duplicates by conventional software methods but did not match with the CRC tags. In the end these were determined not to be duplicate stations.

Match Type E: 13 stations (< 0.2%) were matched as near duplicates by conventional software methods but did not match with the CRC tags. In the end these were determined not to be duplicate stations.

Match Type Z: 4,405 stations (5.7%) were received from SEAS that were unmatched to data received from ISDM. It was determined that ISDM received no real-time data for these stations.

In all cases where duplicate profiles were not identified using the CRC tag, there are good reasons for this. In every case, it has been software or handling error that caused the problem, not something related to the CRC calculation or implementation scheme.

The results above illustrate some of the variations that turn up when comparing data from two sources. The CRC calculation did match about 93% of stations received from SEAS and ISDM. There were 116 stations (< 0.2%) of Match Types D and E received from ISDM, which were

identified to be exact or near duplicates by conventional software matching techniques, but which were determined not to be duplicate stations as their CRC did not match. For various reasons, a proportion of the RTAM which should have been received in delayed mode (from the SEAS program directly) were never sent in real-time and so there were no matching real-time records (Match Type Z). As well, through some reprocessing, the same data arrived twice in delayed mode, though it was only sent once in real-time.

VI. Conclusion

The CRC results to date are very satisfactory and encouraging. It can be said that the CRC tag has performed without flaw. No profiles were identified through the CRC tag as duplicates that were not, and no profiles that were duplicates were identified by the CRC tag as being different. Of interest is that if conventional software methods were relied on, 116 stations would have been identified as duplicates when they were not.

At this point in time (2012), Australia and France have implemented all components as needed. There are several problems still present in the creation and submission of data with the CRC value and possibly some problems still in the match process. So, the only comparisons possible are based on SEAS data.

The GTSPP will continue to monitor the usefulness of the CRC value to provide a unique tag. It is clear, though, that care must be taken to ensure software and processing procedures are carefully carried out.

ACKNOWLEDGEMENTS

Many agencies have played important roles in the development of the GTSPP system. Each participating agency carries out a number of functions in handling the data for the GTSPP. The most important contributors are the collectors of the original data. Without their efforts, this compilation of data and information would not have been possible. The author of this Guide, Dr Charles Sun, would also like to thank Mr Bob Keeley for his leadership in the developing the strategic framework for the CRC algorithm, Mr Mathieu Quellet for CRC calculation at ISDM, Mr Paul Chinn for implementing the CRC algorithm and contribution of Appendix A and figures 2 – 4 of this Guide, Dr Norm Hall and Ms Melanie Hamilton for monitoring and evaluating the algorithm, and Dr Ann Thresher for proofreading this paper. The first draft of the paper was prepared by ISDM. This work was partially supported by the NODC management.

REFERENCES

[1] RAD Data Communications, 1994: “Cyclic Redundancy Check (CRC)” Date Access: 24 August 2011. http://www2.rad.com/networks/1994/err_con/crc.htm

[2] United Nations Educational, Scientific and Cultural Organization (UNESCO), 1999: “Guide to Operational Procedures for the Collection and Exchange of IGOSS Data”, 3rd Revised Edition, Publ. Online. 24 August 2011. Retrieved from http://www.jcommops.org/soopip/mg3.html

APPENDIX A: How to compute the CRC on a WMO FM63 Coded Profile

1. To compute the CRC for the profile in Figure 2, create tokens by breaking the string into individual character groups eliminating all white space, control characters and the equals sign. In the main routine, each token is added to a vector of strings as depicted in Figure 3.

2. Breaking the string into tokens ensures we are only dealing with the data in the bulletin without regard for how the bulletin may have been modified in transit. For example, if a bulletin was created on a Windows computer and rewritten on a UNIX computer, there is the potential of bulletin alteration due to the control characters.

3. Reassemble the bulletin by appending to an empty string the first token after having inserted a space character before it. Continue to append the remaining tokens to the string inserting a space before each token. After appending the last token, append an equal (=) sign to the string. Do not prepend a space character here. The equal sign should be adjacent to the call sign.

4. Find the 8888 group (or 888 in the case of a TESAC) in the string and take all characters following this group as shown in Figure 4 beginning with the space following 88887 token. Invoke the calc_crc32 routine with this string and the string length which for this profile is 2755 characters. The calc_crc32 returns an unsigned long with the hexadecimal value of 5F3BA9DB.

JJVV 15021 2358/ 535015 010735 88887 05206 02205 04205 06205
08205 10205 12206 14206 16206 18206 20206 22206 24206 26206
28206 30206 32206 34206 36205 38204 40201 42199 44196 46195
48194 50193 52190 54185 56182 58181 60180 62177 64175 66173
68172 70171 72170 74169 76168 78166 80165 82164 84163 86161
88159 90155 92154 94154 96154 98153 99901 00153 02152 04151
06151 08150 10150 12150 14150 16150 18150 20149 22149 24149
26148 28148 30148 32148 34148 36147 38147 40147 42146 44146
46146 48145 50145 52144 54144 56143 58143 60143 62143 64142
66142 68142 70142 72142 74142 76141 78141 80141 82141 84140
86140 88140 90140 92139 94139 96139 98139 99902 00139 02139
04138 06138 08138 10137 12137 14137 16136 18136 20135 22135
24135 26134 28134 30134 32134 34134 36133 38133 40132 42132
44132 46131 48131 50130 52130 54130 56130 58130 60129 62129
64129 66129 68129 70128 72128 74128 76128 78128 80128 82127
84127 86126 88126 90126 92125 94125 96125 98125 99903 00125
02125 04124 06124 08124 10124 12124 14124 16124 18124 20124
22124 24124 26123 28123 30123 32123 34123 36122 38122 40122
42122 44121 46121 48120 50120 52120 54119 56119 58119 60118
62118 64117 66117 68117 70117 72117 74117 76116 78116 80116
82115 84115 86115 88115 90115 92114 94114 96114 98114 99904
00114 02113 04113 06112 08112 10112 12112 14112 16111 18111
20111 22111 24110 26110 28110 30109 32109 34108 36108 38107
40107 42106 44106 46105 48105 50105 52104 54104 56103 58103
60102 62102 64101 66101 68101 70101 72101 74101 76100 78099
80098 82098 84097 86097 88096 90096 92096 94095 96095 98094
99905 00094 02094 04094 06094 08094 10093 12093 14092 16091
18090 20090 22090 24090 26090 28089 30089 32089 34089 36089
38088 40088 42087 44087 46086 48085 50085 52085 54084 56084
58083 60083 62082 64082 66081 68081 70081 72081 74080 76079
78079 80079 82078 84078 86078 88078 90077 92076 94076 96075
98074 99906 00074 02073 04073 06073 08073 10073 12072 14072
16072 18071 20071 22080 24096 26097 28091 30074 32081 34091
36091 38090 40090 42089 44088 46088 48088 50088 52088 54088
56089 58089 60089 62088 64087 66087 68088 70090 72095 74109
76113 78119 80120 82120 84122 86123 88123 90124 92126 94126
96126 98126 99907 00123 02127 04127 06128 08129 10128 12128
14128 16128 18129 20132 22132 24133 26133 28134 30135 32135
34134 36134 38134 40134 42133 44133 46133 48132 50132 52132
54132 56131 58131 60130 62130 64129 66129 68128 70128 72128
74127 76126 78125 80125 82124 84123 86121 88121 90120 92119
94119 96118 98117 99908 00116 02115 04114 06113 08112 10110
12109 14108 16107 18106 20105 22104 24103 26101 28100 30099
32098 34097 36097 38096 40096 42096 44095 46094 48093 50092
52091 54091 56090 58089 60089 62089 64088 66087 68086 70086
72086 74085 76085 78084 80084 82084 84084 86083 88083 90083
92083 94086 96086 98086 A8RW3=

Figure 2 WMO FM63 coded profile (Courtesy of Paul Chinn)

JJVV
15021
2358/
535015
010735
88887
05206
02205
04205
06205
08205
10205
12206
14206
16206
18206
20206
22206
24206
26206
28206
30206
32206
34206
36205

Figure 3 Vector of tokens from the profile (Courtesy of Paul Chinn)

```
JJVV 15021 2358/ 535015 010735 88887 05206 02205 04205 06205
08205 10205 12206 14206 16206 18206 20206 22206 24206 26206
28206 30206 32206 34206 36205 38204 40201 42199 44196 46195
48194 50193 52190 54185 56182 58181 60180 62177 64175 66173
68172 70171 72170 74169 76168 78166 80165 82164 84163 86161
88159 90155 92154 94154 96154 98153 99901 00153 02152 04151
06151 08150 10150 12150 14150 16150 18150 20149 22149 24149
26148 28148 30148 32148 34148 36147 38147 40147 42146 44146
46146 48145 50145 52144 54144 56143 58143 60143 62143 64142
66142 68142 70142 72142 74142 76141 78141 80141 82141 84140
86140 88140 90140 92139 94139 96139 98139 99902 00139 02139
04138 06138 08138 10137 12137 14137 16136 18136 20135 22135
24135 26134 28134 30134 32134 34134 36133 38133 40132 42132
44132 46131 48131 50130 52130 54130 56130 58130 60129 62129
64129 66129 68129 70128 72128 74128 76128 78128 80128 82127
84127 86126 88126 90126 92125 94125 96125 98125 99903 00125
02125 04124 06124 08124 10124 12124 14124 16124 18124 20124
22124 24124 26123 28123 30123 32123 34123 36122 38122 40122
42122 44121 46121 48120 50120 52120 54119 56119 58119 60118
62118 64117 66117 68117 70117 72117 74117 76116 78116 80116
82115 84115 86115 88115 90115 92114 94114 96114 98114 99904
00114 02113 04113 06112 08112 10112 12112 14112 16111 18111
20111 22111 24110 26110 28110 30109 32109 34108 36108 38107
40107 42106 44106 46105 48105 50105 52104 54104 56103 58103
60102 62102 64101 66101 68101 70101 72101 74101 76100 78099
80098 82098 84097 86097 88096 90096 92096 94095 96095 98094
99905 00094 02094 04094 06094 08094 10093 12093 14092 16091
18090 20090 22090 24090 26090 28089 30089 32089 34089 36089
38088 40088 42087 44087 46086 48085 50085 52085 54084 56084
58083 60083 62082 64082 66081 68081 70081 72081 74080 76079
78079 80079 82078 84078 86078 88078 90077 92076 94076 96075
98074 99906 00074 02073 04073 06073 08073 10073 12072 14072
16072 18071 20071 22080 24096 26097 28091 30074 32081 34091
36091 38090 40090 42089 44088 46088 48088 50088 52088 54088
56089 58089 60089 62088 64087 66087 68088 70090 72095 74109
76113 78119 80120 82120 84122 86123 88123 90124 92126 94126
96126 98126 99907 00123 02127 04127 06128 08129 10128 12128
14128 16128 18129 20132 22132 24133 26133 28134 30135 32135
34134 36134 38134 40134 42133 44133 46133 48132 50132 52132
54132 56131 58131 60130 62130 64129 66129 68128 70128 72128
74127 76126 78125 80125 82124 84123 86121 88121 90120 92119
94119 96118 98117 99908 00116 02115 04114 06113 08112 10110
12109 14108 16107 18106 20105 22104 24103 26101 28100 30099
32098 34097 36097 38096 40096 42096 44095 46094 48093 50092
52091 54091 56090 58089 60089 62089 64088 66087 68086 70086
72086 74085 76085 78084 80084 82084 84084 86083 88083 90083
92083 94086 96086 98086 /8RVG=
```

Figure 4 WMO FM63 coded profile (Courtesy of Paul Chinn)