GRDC Specification

GRDC Near Real-Time Data Format Version 3.0

Abstract:

This document specifies the GRDC Near Real-Time Data Format Version 3.0, which is intended for supporting the exchange of water level and river discharge data. Furthermore, it describes the guidelines that were followed during the format's development and the benefits of its application.

Version	Date	Authors	Remark
1.0	2006-07-31	Maik Bunschkowski	This is the initial version of this document.
1.1	2006-09-20	Maik Bunschkowski	This is the version after the 1st ETN-R Provider Workshop with backwater influence added.

The Global Runoff Data Centre, GRDC operates under the auspices of the World Meteorological Organization (WMO) with the support of the Federal Republic of Germany within the Federal Institute of Hydrology (BfG)

1. Content

2. The task

The Global Runoff Data Centre (GRDC) is the digital world-wide repository of discharge data and associated metadata[1]. It operates under the auspices of the World Meteorological Organization (WMO) within the premises of the Federal Institute of Hydrology (BfG)[2], Koblenz, Germany. As a contribution to the Global Terrestrial Network Hydrology (GTN-H)[3] and the implementation plan for the Global Observing System for Climate (GCOS)[4], GRDC builds up the Global Terrestrial Network for River Discharge (GTN-R). The basic idea of the GTN-R project is to draw together the already available heterogeneous information on near-real time river discharge data provided by individual National Hydrological Services (NHS)[5] and redistribute it in a harmonised way. On behalf of the Joint Research Centre of the EU (EU JRC) GRDC develops the European terrestrial network for River Discharge (ETN-R) as an automated data collection service for near real -time river gauging data, urgently needed for improving the European Flood Alert System (EFAS), a project of the JRC that aims at providing early flood alerts to the NHS.

Within these fields of application, the GRDC wishes to simplify the way, real-time discharge data is exchanged transnationally. The exchange of real-time data in files needs a data format that can be interpreted by both human beings and computers without problems. It has also to be taken into consideration that an import into current standard software packages should be as easy as possible.

With respect to these constraints and by following the additional guidelines from appendix 4.3 a data file format for the exchange of real-time discharge has been defined. Its detailed specification is given in the following sections.

3. Specification of the GRDC near real-time data format version 3.0

Real-time discharge data, as it will be transported by the GRDC data exchange file format, is nearly unstructured. In most observed cases it contains identifiers of the respective gauging stations, time and date of the measurement, water level, river discharge and a number of logical values (flags). The latter are responsible for annotating conditions, which may lead to errors or non-plausible data. In general no station metadata (except the station identifier) will be transported during real-time discharge data exchange.

[1] See: http://grdc.bafg.de/servlet/is/2377/

[2] See: http://grdc.bafg.de/servlet/is/979/

[3] See: http://grdc.bafg.de/servlet/is/1900/

[4] See: http://grdc.bafg.de/servlet/is/2470/

[5] See: http://grdc.bafg.de/servlet/is/1838/

A common approach to real-time discharge data exchange is the following one: the file format definition is built upon text files and the respective datasets are written in separate rows with the values ordered in columns that are separated by using a delimiter character. This definition is supported by a number of current software applications (such as Microsoft Excel) for data import and data export. It is known as the comma-separated values (CSV) file format.

The only differences of the GRDC near real-time data format version 3.0 (in the following also *the format*) to commonly agreed definitions are that

- it does not allow for character escaping,
- it does not allow other characters, than are defined in 7 bit ascii,
- it does not allow the use of the delimiter character inside values and
- it forbids a data record to span multiple lines.

Every file may start with a fixed header (see 3.2), containing information on the contents of the file. The lines of the header are preceded by the crosshatch "#" character. This character may **only** be used in the header of the file. The header information comprises conditions of use, description of the data columns, pointers to necessary additional information and so on. Spacing characters may be blanks or tabulator characters. If they are encountered adjacently to field separators, they are **ignored during import** and **will never be exported**. One line of text contains one entry. Records are separated with CR (0x13) LF (0x10). **Blank lines** will be ignored during import at the GRDC and **will never be exported**.

3.1. Format description of the data records

The following information is necessary for creating or reading files in the GRDC real-time data file format. The datasets inside the data file are written into separate rows and the respective values, which are ordered in columns, are separated by semicolon characters ";". The definition of the allowed value for each column is given in the following.

1. ***National station identifier*** (case insensitive string). (**mandatory**)
2. ***Timestamp of the measurement*** (Format is: YYYY-MM-DD hh:mm:ss) which follows the ISO 8601 and EN 28601 standard. The time must be given as Universal Time Coordinated (UTC) +0. The calculation is up to the provider, who has to consider local daylight saving time also.

   YYYY stands for the 4 digit year (i. e. 2006). MM stands for the two digit month may be from 01 up to 12 (i. e. 06 for June). The part hh:mm:ss stands for the time of the measurement with hh for hours in the 24 hours scheme (00 up to 23), mm for the minutes (00 up to 59) and ss for the seconds (00 up to 59). (**mandatory**)
3. ***Water level*** (number, metres)
4. ***Discharge*** (number, cubic metres per second)
5. ***Missing value*** (logical, 1= is missing / 0= is not missing) (**mandatory**)
   a. Water level
   b. Discharge
6. ***Is value directly determined*** (logical) (1= directly determined/ 0= indirectly determined) (**mandatory**)
   a. Water level
   b. Discharge
7. ***Is data reliable*** (logical, 1= is reliable / 0= is not reliable) (**mandatory**)
   a. Water level
   b. Discharge
8. **Aggregation** (**mandatory**)
   a. **interval** (number, 0..n, if this is equal to 0, no aggregation took place) (in minutes)
   b. **offset** (number, negative or positive offset in minutes of the timestamp to the end of the averaging interval) (**offset can be omitted if interval is 0**)

Definition: Value is 0 if the timestamp of this data set marks the end of the aggregation interval and equal to value 8.a, if the timestamp marks the start of the aggregation interval. It is 0.5 * value 8.a if the timestamp marks the centre of the aggregation interval.

9. ***Is ice cover*** (logical)
10. ***Is ice jam*** (logical)
11. ***Is weedage*** (logical)
12. ***Is influenced by backwater*** (logical)

The values are always given in SI-units (metres, cubic metres per second). The dot "." is used as decimal point. (i. e. "1.8") There must not occur separator chars for big numbers. Empty columns with two consecutive semicolons (";;") are **allowed only**, if the respective **field is not mandatory**. Empty columns with **logical values** will be treated as **false (0)** and empty columns with **numeric values** will be treated as **missing values. Measured zero values** have to be given as **0**.

In case of missing values for numerical fields that are designated by special logical flags (e.g. water level and discharge) we are proposing to use unique values like -999 as optional markers to ensure a better intelligibility for humans.

Starting with the fifth position, flags describing various properties of the data set are given. These are always logical values encoded as integers (0 for false or 1 for true) where true means that the respective condition is met and false that it is not met.

3.2. File header

Every file may start with a fixed header, containing information on the contents of the file. The lines of the header are preceded by the crosshatch "#" character. This character may **only** be used in the header of the file. Header data is neither required nor used by the ETN-R project during import, but will always be exported to make the data files more intelligible for humans. The header lines have a maximum length of 80 characters. Within the header it is allowed to include any textual information like i.e. about the origin of the included datasets (Provider identifier) and timestamp of the file creation. Any information given in the header are **not** used and **not** stored. An example is provided in annex 4.2.

3.3. Naming conventions

The naming conventions described below should be followed for exporting near real-time data in order to prevent accidental data losses and dead locks. The naming conventions will allow for simple versioning of the collected files also.

Files in the GRDC Near Real-Time Data Format Version 3.0 will always be named following this naming convention:

`<Two letter country code>-<providerID>-YYYYMMDDhhmmss-<version>.nrt`

The list of **two letter country codes** (ISO 3166-1) can be found at

`http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1-semic.txt`

and is updated by the ISO, whenever changes to the codes occur.

The **provider identifier** is issued by GRDC, whenever a new provider commits itself to deliver data. It consists of a number larger than 1000.

YYYYMMDDhhmmss means the time of file creation. It is given as a short version of the time format of ISO 8601 and EN 28601 standards. The time must be given in UTC+0. The calculation is up to the provider, who has to consider local daylight saving time also.

YYYY stands for the 4 digit year (i. e. 2006). MM stands for the two digit month and may be from 01 up to 12 (i. e. 06 for June). The part hh:mm:ss stands for the time of the measurement with hh for hours in the 24 hours scheme (00 up to 23), mm for the minutes (00 up to 59) and ss for the seconds (00 up to 59).

Version stands for the version of the GRDC near real-time data format. It is currently 3.0.

.nrt is the expected file name suffix showing that the file contains near real-time data.

4. Appendices

4.1. Example for creating an NRT file from a MySQL database with a single SQL statement

The following SQL statement creates a csv-file named fr-1001-20060727124400-3.0.nrt at the drive c: of the database server on MS Windows XP and follows the format definition from the previous sections. The example has to be adapted to your environment. If you want to transfer the file somewhere else, you could use a shared network directory for data output. The necessary pre-requisites are:

- The system's decimal separator must be set to ".",
- The system's date format must be set in conformance to ISO 8601 and EN 28601. (i. e. 2006-06-21 14:00:00)
- The database's table structure is appropriately set

```
select stat_id, measure_date_time, water_level, discharge, is_missing_value_water_level,
is_missing_value_discharge,  is_value_measured_water_level, is_value_measured_discharge,
is_data_reliable_water_level, is_data_reliable_discharge, aggregation_interval_water_level,
aggregation_offset_water_level, aggregation_interval_discharge, aggregation_offset_discharge,
is_ice_cover, is_ice_jam, is_weedage, is_influenced_by_backwater INTO OUTFILE "c:\fr-1001-
20060727124400-3.0.nrt" FIELDS TERMINATED BY ';' FROM nrt_data;
```

4.2. Example data file content

```
# GRDC-NRT-Format - for the exchange of near real-time hydrological data
# Version: 3.0
# Provider: 1001
# Timestamp: 20060927105359
#
# Disclaimer: Data in this file is provisional and subject to revision.
# Use of the data at your own risk. All times in this file are in UTC +0.
# All information within this header (all lines started with the crosshatch
# character) are only for informational purposes.
#
# The following lines are organised into columns which are separated by
# the semicolon symbol ";". One line of text contains one entry.
# The use of tabulators or blanks before and after values is not allowed.
# All values are in SI units (metres above gauge zero for water level
# and cubic metres per second for discharge). Logical values are always
# encoded as integers (0 for false, 1 for true).
#
# The following list specifies the data fields and their respective columns.
# It is purely informational.
#
# National Station ID;Timestamp;Water level;Discharge;
# is missing value water level;is missing value discharge;
# is directly determined water level?;is directly determined discharge?;
# is data reliable water level?;is data reliable discharge?;
# aggregation interval water level&discharge;
# aggregation offset water level&discharge;
# is ice cover?;is ice jam?;is weedage?;is influenced by backwater?
WSVN 9640018;2006-09-27 00:01:00;5.04;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:02:00;5.04;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:03:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:04:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:05:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:06:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:07:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:08:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:09:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:10:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:11:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:12:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:13:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:14:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:15:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:16:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:17:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:18:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:19:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:20:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:21:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:22:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:23:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:24:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:25:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:26:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
WSVN 9640018;2006-09-27 00:27:00;5.03;0;0;1;1;0;1;0;0;0;0;0;0;0
```

4.3. Guidelines for the development of the NRT data file format for GRDC

The following guidelines were followed during the development of the GRDC near real-time data format version 3.0.

Table 1: Guidelines for the development of a file format for NRT discharge data

Criterion	Description	Importance
Plausible definition	This means an understandable and plausible definition of the provided data format specification. The current definition should allow for an easy extension with new data fields.	**High** (Inconsistent format definitions will easily lead to erroneous data files and inconsistent implementations.)
Effort for using	This means the effort at the customer's side for reading, writing and processing files in the mentioned format.	**High** (The customers should not have to re-develop their applications.)
Definite future	The chosen data file format should allow for the use as real-time file format over a long period and contain the (and only the) data, necessary for NRT data record description.	**High** (The customers should not have to re-develop their applications.)
Lean interface to a library for reading and writing the format	Definition of the most basic functionality (create file, append data record, close file, read next data record, has next data record) only.	**High** (This corresponds with the effort for using the format.)
Understandability of the data format	Will there arise questions from the file format description? Is it necessary to answer the same questions again and again?	**Medium**
Clarity	How difficult is it to understand the structure of the file format and to read files in it for humans? Can file fragments be easily associated to the respective stations?	**Medium** (This is important for changes to files and for finding out, whether or not there are errors. It's also important for easy manual error correction.)
Compatibility	This means the compatibility of the data file format with existing application i. e. GIS-systems, statistic tools, MS Excel, databases, …	**Very high** (GRDC is collecting and providing data to customers. The data has to be usable without writing specialised software tools.