# EUREF Data Flow

## Diagnostics, Proposals and Possible Improvements

T. PANY[1], P. PESEC[1], G.STANGL[2]

## 1. Introduction

The EUREF Permanent Network (EPN) consists of nearly 100 tracking stations (TSs), one global, one regional and six local data centres (DCs) and 12 analysis centres (ACs) (status summer 2000). The number of stations is growing by approximately 20 a year[3]. Because there is no user statistics one can only assume that the number of users will grow considerably in the same way. With the formal acceptance of ETRS89 by many European countries the importance of EPN as the system's representation the monitoring system will also grow in the near future.. All of the components of EPN, including the user segment, are connected by an information flow via internet. The most intrinsic part of it – with respect to amount of bytes and demand of timeliness between the components – consists of the observed data. Users expect a high degree of availability, completeness and quickness from every system, thus EPN has to take into consideration these requirements too. One should not forget that the ACs act also as users within EPN. Thus watching the data flow within EPN, correcting it if required, and developing some general rules seem to be a duty of the EPN staff. Therefore the co-ordinating group of EPN takes care of the data flow too.

## 2. Diagnosis

The Data Flow within EUREF consists of two parts: The uploading of observed data in RINEX format from the TSs to the DCs by the Operational centres (OCs) and the downloading from the DCs by ACs and other users. It is expected that all receiver dependent formats are already converted to RINEX. Usually observation files, navigation files, and meteo files are transmitted. Sometimes status files are added. In addition there is a flow of IGS (e.g. orbits) and EUREF products (e.g. weekly solutions) between ACs and DCs. Some administrative information like log sheets and EUREF-mail will be transmitted between members of EPN. With respect to size and time limits all this traffic is negligible compared to the bulk of observation data. Divided by categories, RINEX files of 24 hours and one hour length are currently transferred. Files are usually transmitted with Hatanaka and unix compression to minimise size and transfer time. However, from the TSs to the Local DCs several variations are used. Files without Hatanaka compression are found for several stations. More unusual are files without unix compression or with other compressions (e.g. ZIP). Concerning the file names capital or small letters are used, sometimes mixed. Unix machines discern between small and capital letters, which usually makes a conversion necessary. The unification should be one of the tasks of the OCs. Unfortunately only two OCs (ASI, GOP) are registered within EUREF, covering about 10% of all stations only. The DCs hold their files in different directory structures using file names either in capital or small letters (except the Z extension of the unix compression). Between DCs and ACs the exchange of data in Hatanaka and unix compressed format is mandatory. There is no information in which form users really want to retrieve the data. However, commercial analysis programs accept RINEX data without compression only, therefore a public decompression procedure will be needed anyway.

As unix compressed files can be decompressed by several packing programs under Windows, only the Hatanaka compression is widely unknown. A sort of Hatanaka decompression mechanism should therefore be installed at the DCs. This can be done by providing separate directories containing RINEX-observation files or by implementing the Hatanaka decompression in the downloading procedure via ftp or http.

### 2.1 24h RINEX Data

More than 95% of the daily RINEX files arrive at the DCs within one day, most of them within few hours after midnight. Concerning the risks of broken communication lines the most critical path lies between TSs and the local DCs. As can be seen at the EUREF homepage the data holding for the same station varies between the DCs too. This may come from late files, where the DCs acting as mirrors do not try to retrieve the data after a couple of days. In principle no DC knows if the special TS has produced data for the time span or not. However, several stations announce the loss of data by EUREF-mail which may be a way to let the DCs know that there is no data. The question is whether the DC should ask the TSs for data which are missing and not declared as outage. A second problem is the filtering of files which are truncated during the transfer.

[1] Thomas Pany, Peter Pesec; Austrian Academy of Sciences, Space Research Institute, Department of Satellite Geodesy; Elisabethstrasse 20, A-8010 Graz, Austria; Tel. +43 316 322607; Fax +43 316 3210914; e-mail Thomas.Pany@oeaw.ac.at, peter.pesec@oeaw.ac.at

[2] Günter Stangl; Federal Office of Metrology and Surveying; Address: Space Research Institute, Department of Satellite Geodesy; Elisabethstrasse 20, A-8010 Graz, Austria; Tel. +43 316 322607; Fax +43 316 3210914; e-mail guenter.stangl@oeaw.ac.at

[3] See figure 3 in C. Bruyninx, M. Becker, G. Stangl, Regional Densification of the IGS in Europe Using the EUREF Permanent GPS Network (EPN), IGS Network Workshop Oslo 2000.

It sometimes happens that files are available at the DCs but cannot be decompressed. A check program which filters out such files is not always used. A response from the DCs to the TSs concerning all those problems is missing.

In summary one can conclude that the transfer via internet for 24h RINEX files is well established. For the EUREF weekly analysis tasks which are carried out about 10 days after the observations the transfer is sufficiently fast. Apart from the minor problem of broken files only local improvements at some stations are needed. The communication between the members of the transfer chain is not sufficient in case of missing or unusable data and has to be improved.

### 2.2 Hourly RINEX Data

Transmittance of hourly RINEX data have been promoted because there is a need for a near real-time analysis, especially in numerical weather prediction. Shortly after the end of each hour RINEX files should be sent from the TS to the corresponding DC. The transfer is checked by a flag file attached to transmitted data package. If the flag file is received correctly all foregoing files are assumed to have been transferred correctly too. In total the amount of traffic in size and time increases, but the risks of breaks and waiting times decrease considerably because smaller file sizes are involved. The number of stations which are able to send hourly data increases steadily within EUREF. Obviously, all this transfers have to been automated. It can be seen from the protocols however, that not all TS try to repeat the transfer when it failed. A second problem is the arrival of data from several TSs at the DCs within the same minute which causes capacity problems, a time sliced "get" action of the DCs may be an alternative. Since the 24h RINEX files and the daily sum of hourly files should contain the same information it was proposed to cancel the 24h files and reconstruct them from the daily files. Looking at the protocol file CHECK_HOURLY.BKG missing hourly files can be seen. During June 2000 one to five of a total of 20 stations miss at least one hour per day. If one concatenates hourly files gaps would be introduced, coming mainly from unsuccessful data transfer. A communication procedure for missing files between TSs and DCs does not exist officially. The third problem is the "late" (15 minutes or more after closing the file) arrival of some stations because a near real-time service like weather prediction needs an output as fast as possible to start with its predictions. Up to now (June 2000) no time schedule for hourly transfers exists.

In total the transfer of hourly files is working well and is able to serve the demands of near real-time services. But it is not yet sufficiently organised to replace the daily transfer of 24h files as a whole.

## 3. Proposals

The following proposals are intended to work within EUREF, proposals concerning a possible change of the boundary conditions have been omitted. The activities which can be done within EUREF have only marginal influence on the surrounding facilities. It would be nice to have a transmission medium for its own which is fast, reliable and cheap. But, living in real world with limited finances one has to accept that the internet is the only one which meets the requirements of EPN. Thus one has to live with connection problems and to react to those problems. There is only a marginal chance to find a system which downloads all GPS receivers with a uniform data structure, edits all needed information correctly and transfers the data automatically without any problems. However, there is a chance to fulfil the requirements to do so.

### 3.1 Fall-Back Strategies

Missing data are by far the most crucial problem. If a TS cannot provide data due to operational problems generally there is no way to get around it. If a receiver fails one can think about strategies of borrowing a receiver from a receiver pool established by the EUREF community, but in most cases things are more complex and the problems are solved quicker than installing a spare receiver. However, if data are missing due to broken transfer lines different ways can be developed to make the data available. Some proposals in this case are:

- Repeated trials by ftp at several epochs. This is the standard procedure nowadays either by manual or automatic processes.
- Cutting daily files into hourly pieces and sending all files to the Local DC. An announcement is required, because the DC usually does not concatenate those files.
- Sending the file as an attachment of an e-mail to the DC is a way to overcome the time-critical ftp procedure. The disadvantage is that there is no good automatic algorithm available to retrieve binary data files from e-mails.
- Installation of a direct modem connection to the DC server makes a TS independent from an internet provider and blocked internet services. This way is costly for both, TS and DC, apart from security problems. It should only be considered for remote stations.

If the data cannot be provided according to a failure of the corresponding DC the fall-back strategies of Figure 1 are proposed. A TS has to try at least another DC or the RDC directly. It would be a major disaster not only for the EPN, but for the whole net community if none of the DCs would be reachable, because the DCs are already well distributed over Europe.

Therefore a general directive for the TS can be given like:

- Try to reach your fall-back DC,
- Try to reach the Regional DC or its mirror (see below),
- Try all other DCs.

In principle the same procedure should hold for the DCs. There is one exception, because up to now only one Regional DC (BKG) exists. Unfortunately it is on the top of the hierarchy and holds most of the collected data. If this DC experiences an outage data collection by users would become uncomfortable and incomplete, because the Local DCs do not hold all data, because some TSs send their data directly to BKG. Therefore a mirror site is proposed as shown in Figure 1. This mirror holds the data of BKG and can be accessed in case of an eventual outage. In principle one could propose a distributed storage network, established by the Local DCs too, but this would be less comfortable for users. In that case users would have to know the complete data-holding list of all local DCs. If a mirror has been established, the fall-back strategies for all DCs would be the same, namely to direct their data flow to the mirror. Otherwise each DC has to be given a separate instruction, which station would be sent to which DC.

The overall question is when the fall-back rules should be applied. Considering the possible delays of daily files a lag of one day seems to be acceptable. That would mean that, if a DC cannot be reached within 24 hours, the data should be transmitted to the fall-back DC. After an official announcement that the missing DC is available again the data flow should be switched back within 24 hours.

### 3.2 Consistency Checks

Providing consistent and usable data is the major task of the DCs. Whereas the TSs are responsible for the contents of the data files these contents have to be checked for usability. The checks should be restricted to formal criteria at the time of transmission, the results of the ACs can be used for checking the data quality afterwards. Formal checks concentrate on the RINEX header and optionally on the first line of each RINEX data cluster. Because the file name contains the date implicitly, the correct date is also a matter of formal checking. Table 1 gives an overview which parameters of an observation file should be checked. All checks are chosen in compliance with the EUREF regulations. At present the files are checked officially after the data transmission. File checks are done by several DCs with different software and strategies, but there are no common rules. The proposal is to set minimum standards to file checks which are applied at each DC. The original thought was that OCs should do this task independently from the TSs, but this would imply the creation of several OCs first. It seems to be easier to implement the rules at the DCs which partly act as "silent" Ocs. For the automatic checking there are several software candidates which may be combined into a checking package running on several systems. The three most important seem to be the Hatanaka compression/decompression, teqc from UNAVCO and the software from BKG.

The consequence of detecting blunders in files should be that the file has to be redrawn from the database until the problem is solved. The best way would be if such files would not be present at all in any database. Otherwise some non-EUREF DCs may hold inconsistent files which leads to confusion among users.

Tab. 1: Main parameters of RINEX files to be checked within data flow

| Parameter to check | Mainly checked for |
|---|---|
| Compressed file | Correct decompression |
| All header contents | Correct position and length |
| MARKER NAME | Characters 1-4 (file name, EUREF log sheet) |
| MARKER NUMBER | Domes number (EUREF log sheet) |
| REC TYPE | Receiver type EUREF (log sheet) and IGS (existence) |
| ANT TYPE | Antenna type EUREF (log sheet) and IGS (existence) |
| DELTA H/E/N | Antenna eccentricities (EUREF log sheet) |
| TYPES OF OBSERV | Phase dual frequency, code at least one |
| TIME OF FIRST OBS | Correct date (file name) |
| END OF HEADER | Existence |
| Date of epoch | Correct date (file name) |

### 3.3 Communication

Apart from the regular data flow an improved flow of information among the members of the EPN is required. Many of the problems can be solved in a bilateral way, e.g. between one TS and one DC. However, some problems are of major concern which may require more public communication. This could be:

- Announcement of data loss of more than 12 hours in daily files via EUREF-mail
- Announcement of resubmission of daily data files via EUREF-mail
- Automatic failure procedure for hourly files at TSs within 24 hours
- Feed-back from DCs to TSs concerning inconsistent files
- Announcement of outages of TSs and DCs longer than 24 hours via EUREF-mail

Concerning daily files a growing number of TSs announces their observation outages or the resubmission of files via EUREF-mail. As mentioned above the official file check will be done at the end of the transfer chain nowadays. A faster feed back by the first DC receiving the inconsistent file should be preferred.

The most difficult task is to establish an automatic communication about missing hourly files. At first a proper balance has to be found between the needs of fast data delivery requested by the users and the amount of effort a TS has to invest to cover all problems. It is anticipated that a download failure leading to a maximal delay of three hours should be tolerated if the respective files actually exist at the TS. A "non-existing" information would be advantageous. Consecutive missing hourly files should be a matter of concern. During daytime at least a message to the responsible person should be sent or from him/her a reaction should be set. Every DC is obliged to have a summary log file <check_hourly.x> marking the existence of all files available at the database. This can be extended to the TSs, together with a small watch-dog which reports missing files to the administrator.

The decision when an outage needs a switch of a data flow should be made by the Data Flow Coordinator of EPN. A personal contact with the missing TS or DC before the decision is advisable

## 4. Improvements and Conclusions

As usual, the target is to become faster, more reliable and more comfortable. Otherwise one has to trade the efforts against the gain. As an example it has to be investigated whether a gain of five minutes for sending the hourly files earlier is of actual importance. At first all computers have to be synchronised (Tardis ?) because computer clocks can differ by this amount running without time steering. Secondly, one computer which downloads several stations by modem is dependent on the quality of the communication line and would risk a break which may result in a complete blocking of all download procedures.

The positive aspects are that by introducing rules and cheap software one can achieve improvements:

- Despite of the problems mentioned above a faster transfer of hourly files would reduce the latency time thus being an improvement. Time slices can be attributed to each of the stations depending on their special needs, thus enabling more secure data transmission and easier control.
- The overall availability of the files will be improved by setting the fall-back strategies already proposed. The cost of establishing a mirror reduces considerably if a still existing DC would take over this task.
- Checking the data consistency improves the reliability of products "made in EUREF", which comprises also the raw data.
- Because most of the DCs use already checks one has only to standardise the checks and make them public.
- Communication is a time consuming procedure, but necessary for each co-operative work like the data flow of EUREF. However, communication reduces if problems are vanishing. In this case more communication between TS and DC for solving problems will reduce communication efforts between DCs and users considerably.

The function of a Data Flow Coordinator will help to observe the general flow and may improve the reactions in case of broken transfer chains. Apart from some regulations the daily work of all members of EPN will do the main task of improving the EUREF data flow.

Fig. 1: Current data flow for RINEX 24h and hourly files and planned fall back strategies