Chapter 10

Seismic Data Formats, Archival and Exchange

(Version August 2011; DOI: 10.2312/GFZ.NMSOP-2_ch10)

Bernard Dost[1], Jan Zednik[2], Jens Havskov[3],
Raymond J. Willemann[4], and Peter Bormann[5]

[1] Seismology Division KNMI, P.O. Box 201, 3730 AE De Bilt, The Netherlands; E-mail: Bernard.Dost@knmi.nl
[2] Geophysical Institute AS CR, Bocni II/1401, 141 31 Prague 4, Czech Republic; E-mail: jzd@ig.cas.cz
[3] University of Bergen, Department of Earth Science, Allégaten 41, N-5007 Bergen, Norway; E-mail: Jens.Havskov@geo.uib.no; Lars.Ottemoller@geo.uib.no;
[4] IRIS Headquarters, 1200 New York Ave., Suite 800, Washington, DC 2005; E-mail: ray@iris.edu
[5] Formerly GFZ German Research Center for Geosciences, Telegrafenberg, 14473 Potsdam, Germany; E-mail: pb65@gmx.net.

page

10.1 Introduction

Seismology entirely depends on co-operation, both national and international cooperation. Only the accumulation of large sets of compatible high quality data in standardized formats from many stations and networks around the globe and over long periods of time will yield sufficiently reliable long-term results in event localization, seismicity rate and hazard assessment, investigations into the structure and rheology of the Earth interior and other priority tasks in seismological research and applications.

For almost a century, only parameter readings taken from seismograms were exchanged with other stations and regularly transferred to national or international data centers for further processing. Because of the uniqueness of traditional paper seismograms and lacking opportunities for producing high-quality copies at low cost, original analog waveform data, cumbersome to handle and prone to damage or even loss, were rarely exchanged. The procedures for carefully processing, handling, annotating and storing such records have been extensively described in the 1979 edition of the Manual of Seismological Observatory Practice (Willmore, 1979) in the chapter *Station operation*. They are not repeated here. Also the traditional way of reporting parameter readings from seismograms to international data centers such as the U.S. Geological Survey National Earthquake Information Center (NEIC), the International Seismological Centre (ISC) or the European Mediterranean Seismological Centre (EMSC) are outlined in the old manual in detail in the section *Reporting output*. They have not changed essentially since then, although methods to exchange parameter data did change dramatically. On the other hand, respective working groups on parameter formats of the IASPEI and of its regional European Seismological Commission (ESC) have meanwhile debated for many years, how to make these formats more homogeneous, consistent and flexible so as to better accommodate also other seismologically relevant parameter information.

Any data report must, of course, follow a format known to the recipient in order to be successfully parsed. Some of the goals for any format are:

- *concise* avoiding unnecessary expense in transmission and storage
- *complete* providing all of the information required to use the data
- *transparent* easily read by a person, perhaps without documentation
- *simple* straightforward to write and parse with computer programs

Traditional formats for reporting parameter data sacrificed simplicity, transparency and even sometimes completeness in favor of the other goals. With the falling cost of data storage and exchange, modern formats more often sacrifice conciseness in favor of transparency and simplicity.

Modern parameter formats, in addition, are usually extensible and include “metadata”. An extensible format includes some way for new types of data to be introduced without either collecting all of the new information into unformatted comment strings or making messages with the new data types unreadable by old parsers. “Metadata” are information about the data, such as how and by whom the data were prepared. For waveform data the term metadata includes all information on the station and sensor/data acquisition system that produces the data.

The Telegraphic Format (TF), as documented in the Manual of Seismic Observatory Practice (Willmore, 1979) is an extreme example of a traditional format for reporting and exchanging parameter data. Since telex was very expensive compared with the modern communication costs, conciseness was the paramount goal even to the point of occasional ambiguity. The year of the data, for example, might be excluded if the recipient could probably infer it. The format was intended for use in an era when many stations were isolated and could report little more than their own phase readings, so event parameters such as hypocenter and magnitude are relegated to a secondary role. The TF incorporated further restrictions due to the special limitations of telex messages, such as no lower-case letters and sometimes no control over line breaks.

A seismic network with modern, calibrated instruments can provide far more information than telegraphic format allows, while low-cost e-mail has eliminated the restrictions and high costs of telex messages. Consequently, since at least 1990 most seismic parameter data have been stored and exchanged in modern formats that are more complete, simpler and usually more transparent than the Telegraphic Format. But until recently there was no generally accepted standard modern format. A major step forward in this direction was made by the Group of Scientific Experts (GSE) organized by the United Nations Conference on Disarmament. It developed GSE/IMS formats (see 10.2.4) for exchanging parametric seismological data in tests of monitoring the Comprehensive Test Ban Treaty (CTBT) (see 10.2.4) which became popular also with other user groups. Seismological research, however, has a broader scope than the International Monitoring System (IMS) for the CTBT. Therefore, a new *IASPEI Seismic Format* (ISF), compatible with the IMS format but with essential extensions, has been developed and adopted by the Commission on Seismological Observation and Interpretation of the International Association of Seismology and Physics of the Earth´s Interior at its meeting in Hanoi, August 2001. It is the conclusion of a 16-year process seeking consensus on a new format and fully exploits the much greater flexibility and potential of E-mail and Internet information exchange as compared to the older telegraphic reports (see 10.2.5). Since 2001 new developments have been introduced due to technology changes. The introduction of XML resulted in the requirement of a standard schema for the representation of parameter data: QuakeML (https://quake.ethz.ch/quakeml/QuakeML). QuakeML is a collaborative effort, which started within the NERIES project, and presents an open standard.

Digital waveform data, however, are nowadays by far the largest volume of seismic data stored and exchanged world-wide. The number of formats in existence and their complexity far exceeds the variability for parameter data. With the wide availability of continuous digital waveform data and unique communication technologies for world-wide transfer of such complete original data, their reliable exchange and archival has gained tremendous importance. Several standards for exchange and archival have been proposed, yet a much larger number of formats are in daily use. The purpose of the section on digital waveform data is to describe the international standards and to summarize the most often used formats. In addition, there will be a description of some of the more common conversion programs.

Due to the rapid increase of seismological stations, the practice of an international registry of station names (ISC, NEIC) became difficult to maintain. New definitions of station codes have been discussed within IASPEI and FDSN since 2005 and approved by IASPEI and its Commission on Seismic Observation and Interpretation (CoSOI) at the 2009 Cape Town General Assembly (see IS 10.3).

However, beforehand, a short description of the most common parameter formats will be given in the following section.

10.2 Parameter formats

Parameter formats deal with all earthquake parameters like origin times, hypocenter coordinates (longitude, latitude and depth), phase names and phase arrival times, maximum amplitudes and periods of different seismic phases or wave groups as recorded by different types of seismographs, magnitudes calculated from such amplitude and period readings, etc. Until recently, there were no real standards, except the Telegraphic Format (TF) used for many years to report phase arrival, amplitude and period data to international agencies (Willmore, 1979; chapter "Reporting output"). The format is not used for processing. There have been attempts to modernize TF for many years through the IASPEI Commission of Practice (now the Commission on Seismological Observation and Interpretation) and as mentioned in the introduction, the IASPEI Seismic Format (ISF) was approved as a standard in 2001 (see section 10.2.5). Recently IASPEI (2005 and 2011) has adopted measurement standards for several widely used magnitudes. The standard nomenclature for reporting such magnitudes and related amplitude and period data in accordance with the ISF format have been outlined in IS 3.3 as well as in Chapter 3, sections 3.2.7.3 and 3.2.7.5. An example is given in Tab. 10.5 below. In practice, however, still many different formats are used and the most dominant ones have come from popular processing systems. In the following, some of the most well known formats will be briefly described. For complete description of the formats, the reader is referred to original manuals or publications.

10.2.1 HYPO71

The very popular location program HYPO71 (Lee and Lahr, 1975) has been around for many years and has been the most used program for local earthquakes. The format was therefore limited to work with only a few of the important parameters. Tab. 10.1 gives an example.

Tab. 10.1 Example of an input file in HYPO71 format. Each line contains, from left to right: Station code (max 4 characters), E (emergent) or I (impulsive) for onset clarity, polarity (C – compression; D – dilatation), year, month, day, and time (hours, minutes, seconds, hundredth of seconds) for P-Phase, second for S-phase (seconds and hundredth of seconds only), S-phase onset and, in the last column, duration. The blank space between ES and duration has been used for different purposes like amplitude. The last line is a separator line between events and contains control information.

```
FOO EPC  96 6 6 64848.47       62.67ES                                  136
MOL EPC  96 6 6 64849.97       65.87ES                                  144
HYA EP   96 6 6 64856.78       78.07ES                                  135
ASK EP   96 6 6 649 2.94       34.72ES                                  183
BER EPC  96 6 6 649 7.56       36.61ES
EGD EPD  96 6 6 649 5.76       40.53ES
                  10  5.0
```

The format is rather limited since only P or S phase names can be used and the S-phase is referenced to the same hour-minute as the P-phase and the format cannot be used with teleseismic data. However, the format is probably one of the most popular formats ever for local earthquakes. The HYPO71 program has seen many modifications and the format exists in many forms with small changes.

10.2.2 HYPOINVERSE

Following the popularity of HYPO71, several other popular location programs followed like Hypoinverse (Klein, 1978) and Hypoellipse (Lahr, 1989). Tab. 10.2 gives an example of the input format for Hypoinverse.

Tab. 10. 2 Example of the Hypoinverse input format. Note that year, month, day, hour, min is only given in the header and only one phase is given per line.

```
96 6 60648
FOO EPC  48.5 136
FOO ES   62.7
MOL EPC  50.0 144
MOL EPC  50.9
MOL ES   65.9
```

10.2.3 Nordic format

In the 1980’s, this was one of the first attempts to create a more complete format for data exchange and processing. The initiative came from the need to exchange and store data in Nordic countries and the so-called Nordic format was agreed upon among the 5 Nordic countries. The format later became the standard format used in the SEISAN data base and processing system and is now widely used. The format tried to address some of the shortcomings in HYPO71 format by being able to store nearly all parameters used, having space for extensions and useful for both input and output. An example is given in Tab. 10.3.

Tab. 10. 3 Example of Nordic format. The data is the same as seen in Tabs. 10.1 and 10.2. The format starts with a series of header lines with type of line indicated in the last column (80) and the phase lines are following the header lines with no line type indicator. There can be any number of header lines including comment lines. The first line gives among other things, origin time, location and magnitudes, the second line is the error estimate, the third line is the name of the corresponding waveform file and the fourth line is the explanation line for the phases (type 7). The abbreviations are: STAT: Station code, SP: component, I: I or E, PHAS. Phase, W: Weight, D: polarity, HRMM SECON: time, CODA: Duration, AMPLIT: Amplitude, PERI: Period, AZIMU: Azimuth at station, VELO: Apparent velocity, SNR: Signal-to-noise ratio, AR: Azimuth residual of location, TRES: Travel time residual, W: Weight in location, DIS: Epicentral distance in km and CAZ: Azimuth from event to station.

```
1996  6 6 0648 30.4 L  62.635   5.047 15.0  TES 13 1.4 3.0CTES 2.9LTES 3.0LNAO1
 GAP=267        5.92      18.8    43.0 31.8 -0.5630E+03  0.8720E+03 -0.3916E+03E
 1996-06-06-0647-46S.TEST__011                                                 6
 STAT SP IPHASW D HRMM SECON CODA AMPLIT PERI AZIMU VELO SNR AR TRES W  DIS CAZ7
 FOO  SZ EP     C  648 48.47  136                               -0.110  116 180
 FOO  SZ ESG       649  2.67                                     0.710  116 180
 FOO  SZ E         649  2.89       426.4  0.3                           116 180
 MOL  SZ EP     C  648 49.97  144                               -0.310  129  92
 MOL  SZ EPG    C  648 50.90                                     0.410  129  92
 MOL  AZ E         649  5.86                                            129  92
 MOL  SZ ESG       649  5.87                                     0.410  129  92
 MOL  SZ E         649  6.98       328.6  0.6                           129  92
 HYA  SZ EP        648 56.78  135                                0.810  174 159
 HYA  SZ IP     D  648 56.78                                     0.810  174 159
 HYA  SZ EPG    D  648 57.56                                     0.110  174 159
 HYA  SZ ESG       649 18.07                                     0.610  174 159
 NRA0 SZ  Pn      0649 24.03                  309.6  8.5 139  5 -0.410  403 119
 NRA0 SZ  Pg      0649 32.60                  305.6  7.285.2  1  0.410  403 119
 NRA0 SZ  Lg      0650 22.05                  302.0  4.016.0 -1 -0.410  403 119
```

10.2.4 The GSE/IMS formats

Also in the late eighties, a new format was created for exchange of data within the International Monitoring System (IMS) of the Comprehensive Test Ban Treaty Organization (CTBTO), formally called the GSE (Group of Scientific Experts) parameter format. The format used in the GSE´s Technical Test 3 (GSETT-3) was designated GSE2.0 and came to be used even by seismologists uninvolved in CTBT monitoring.

Following GSETT-3, a significantly revised format was originally designated GSE2.1, but renamed IMS1.0 when it was adopted for use by the International Data Center planned to monitor the CTBT when it enters into force. The format IMS1.0 is similar in structure to the Nordic format, however more complete in some respects and lacking features in other respects. A major difference is that the line length can be more than 80 characters long, which is not the case for any of the previously described formats. After SEISAN, it is the first major format for which completeness or readability was recognized as a more important design goal than conciseness. The IMS1.0 format is available from the ISC web site via http://www.isc.ac.uk/standards/isf and ftp://ftp.isc.ac.uk/pub/isf/isf_ext.pdf. IMS1.0 has been used extensively for data exchange within the institutions participating in the IMS and also been used for data exchange outside IMS like in the popular AutoDRM system (http://www.seismo.ethz.ch/prod/autodrm/index_EN) and for transmission to international centers. However, IMS has been used less as a processing format than HYPO71 and Nordic formats.

Tab. 10.4 An example of the IMS1.0 parameter format. The example contains the same data as given in Tabs. 10.1 to 10.3. The first lines are message information etc. The remaining lines are more or less self-explanatory. Note that more information, with a higher accuracy, can be given for each phase (like magnitude) than in the Nordic format. On the other hand, information like component and event duration is missing. These are added in the new ISF format.

```
BEGIN GSE2.0
MSG_TYPE DATA
MSG_ID 1900/10/19_1711 ISR_NDC
DATA_TYPE ORIGIN GSE2.0
EVENT 00000001
   Date       Time       Latitude Longitude    Depth    Ndef Nsta Gap    Mag1  N    Mag2  N
rms   OT_Error      Smajor Sminor Az        Err   mdist  Mdist     Err        Err        Err
1996/06/06 06:48:30.4     62.6350    5.0470     15.0      25   13 267             ML 2.9  8
1.40   +-  5.92       0.0    0.0    0    +- 31.8    1.04   4.84              +-0.3
Sta     Dist  EvAz     Phase      Date       Time     TRes  Azim  AzRes  Slow  SRes Def   SNR
Amp   Per   Mag1   Mag2   Arr ID
FOO     1.04 180.0 mc  P       1996/06/06 06:48:48.5  -0.1                          T
FOO     1.04 180.0 m   SG      1996/06/06 06:49:02.7   0.7                          T
FOO     1.04 180.0 m           1996/06/06 06:49:02.9
426.4  0.30 ML 3.2        00000003  (from previous line)
MOL     1.16  92.0 mc  P       1996/06/06 06:48:50.0  -0.3                          T
MOL     1.16  92.0 mc  PG      1996/06/06 06:48:50.9   0.4                          T
MOL     1.16  92.0 m           1996/06/06 06:49:05.9
MOL     1.16  92.0 m   SG      1996/06/06 06:49:05.9   0.4                          T
MOL     1.16  92.0 m           1996/06/06 06:49:07.0
NRA0    3.62 119.0 m   Pn      1996/06/06 06:49:24.0  -0.4 309.6    5.0   8.5       TAS  13.9
(from previous line)
NRA0    3.62 119.0 m   Pg      1996/06/06 06:49:32.6   0.4 305.6    1.0   7.2       TAS  85.2
NRA0    3.62 119.0 m   Lg      1996/06/06 06:50:22.0  -0.4 302.0   -1.0   4.0       TAS  16.0
(from previous line)
STOP
```

10.2.5 The IASPEI Seismic Format (ISF)

The need for a common and widely accepted parameter format for comprehensive seismological data exchange has led to the IASPEI Seismic Format (ISF), adopted as standard in August 2001. ISF conforms to the IMS.1.0 standard but has essential extensions for reporting additional types of data. This allows the contributor to include complementary data considered to be important for seismological research and applications by the IASPEI Commission on Seismological Observation and Interpretation. The format looks almost like the IMS1.0 example in Tab. 10.4 above, except for the extensions. The ISF has been comprehensively tested at the ISC and NEIC and incompatibilities have been eliminated. The detailed description of the ISF is available from the ISC home page and kept up-to-date there (see ftp://ftp.isc.ac.uk/pub/isf/isf_ext.pdf). It is not reproduced in this manual.

Consensus on the ISF was reached partly by including many optional items, so the format is not as simple as some alternatives. Despite this, the completeness, transparency, extensibility and metadata of ISF are expected to make it very widely used. Wide use of ISF will bring back the advantages of a generally accepted standard so that it becomes easier to exchange data, re-use data collected for past projects, and employ programs developed elsewhere.

In the Information Sheets IS 10.1 and IS 10.2, examples are given how event parameter data and unassociated parameter readings by seismic stations are reported according to the IMS format with ISF extensions.

Tab. 10.5 below gives an abridged example of an USGS/NEIC Hydra System data printout in the IMS1.0 format of calculated event parameters and station parameter readings. Yet, for brevity, several columns in the original listing have been left out in order to highlight the new, more differentiated way of magnitude, amplitude and period reporting It helps to recognize easily which amplitudes and magnitudes have been measured according to the IASPEI (2005) recommended standards and which ones not, whether the amplitudes are displacement or velocity amplitudes and at what “amplitude phase” time these parameters have been “picked”. This assures later on an unambiguous reproduction or check of these parameter determinations.

Tab. 10.5 Cut-out of an event and station parameter plot in IMS1.0/ISF parameter format produced by the USGS/NEIC HYDRA automatic location and analysis system. Note the IASPEI (2005) standard magnitudes mb, mb_Lg, mB_BB, Ms_BB and Ms_20. BB stands for unfiltered velocity broadband records, 20 for surface waves with periods of 20±2 s measured in WWSSN-LP filtered records. In contrast, mb and mb_Lg are measured on WWSS-SP filtered records. Letters preceding the magnitude symbol in the "amplitude-phase name" column stand for: I - IASPEI standard, A - displacement amplitude in nm and V - velocity amplitude in nm/s. The next column gives the time at which the respective amplitudes and periods have been measured. Additionally, NEIC also determines several non-standard magnitudes.

```
BEGIN IMS1.0
MSG_TYPE DATA
MSG_ID 30000Z9N_43 HYDRA_ORANGE
DATA_TYPE BULLETIN IMS1.0:SHORT
The following is an UNCHECKED, FULLY AUTOMATIC LOCATION from the USGS/NEIC Hydra
System
Event    15694

   Date       Time        Err   RMS Latitude Longitude  Smaj  Smin  Az Depth
2009/09/29 17:48:13.26   2.94  1.45 -15.5267 -172.0703   6.5   5.8 133  28.4

Magnitude  Err Nsta Author      OrigID
Mb_Lg  6.0 0.5    1 NEIC           0
Ms_VX  8.2 0.1   23 NEIC           0
mb     7.2 0.0  243 NEIC           0
Mwp    7.8 0.0  179 NEIC           0
mB_BB  7.7 0.0  246 NEIC           0
Ms_BB  8.3 0.1  134 NEIC           0
Mwb    7.7 0.0   97 NEIC           0
Ms_20  8.0 0.0  161 NEIC           0
Mwc    8.1 0.0   71 NEIC           0
M      7.9 0.1 1064 NEIC           0
Mwp    7.9 0.0    4 PMR            0

Sta      Dist  EvAz Phase        Time            Amp    Per  Magnitude    ArrID
KNTN   12.68   1.6 IAmb_Lg  17:55:00.090       7785.9  0.98   Mb_Lg  6.0 BHZIU00
KNTN   12.68   1.6 IVMs_BB  17:56:03.398    2169276.1 10.00   Ms_BB  7.7 LHZIU00
TARA   22.39 317.2 P        17:53:11.829                                  BHZIU00
TARA   22.39 317.2 IAmb     17:53:45.055      12080.1  1.25   mb     7.2 BHZIU00
TARA   22.39 317.2 MMwp     17:53:53.279    1728974.9 41.45   Mwp    7.6 BHZIU00
TARA   22.39 317.2 IVmB_BB  17:54:07.329     206688.6  5.40   mB_BB  7.8 BHZIU00
OUZ    23.45 210.6 P        17:53:22.710                                  HHZNZ10
OUZ    23.45 210.6 IAmb     17:53:47.450      11261.4  1.22   mb     7.3 HHZNZ10
OUZ    23.45 210.6 IVmB_BB  17:53:49.880     314772.2  9.74   mB_BB  8.0 HHZNZ10
OUZ    23.45 210.6 IAMs_20  18:01:02.679       6314.2 20.00   Ms_20  8.1 LHZNZ10
OUZ    23.45 210.6 IVMs_BB  18:02:47.679    3821858.4 16.00   Ms_BB  8.4 LHZNZ10
OUZ    23.45 210.6 AMs_VX   18:02:54.449    6326077.5 15.00   Ms_VX  8.3 LHZNZ10
```

10.3 Digital waveform data

Many different formats for digital data are used today in seismology. For a summary and the abbreviations used, see the following sections. Most formats can be grouped into one of the following five classes:

1. Local formats in use at individual stations, networks or used by a particular seismic recorder (e.g., ESSTF, PDR-2, BDSN, GDSN; mini-SEED).
2. Formats used in standard analysis software (e.g., SEISAN, SAC, AH, BDSN; mini-SEED).
3. Formats designed for data exchange and archiving (SEED, GSE).

4. Formats designed for database systems (CSS, SUDS).
5. Formats for real time data transmission (IDC/IMS, Earthworm; mini-SEED).

Use of the term "designed" in describing Class 3 and 4 formats is intentional. It is usually only at this level that very much thought has been given to the subtleties of format structure which result in efficiency, flexibility, and extensibility.

The four classes (1-4) show a hierarchical structure. Class 4 forms a superset of the others, meaning that classes 1-3 can be deduced from it. The same argument applies to class 3 with respect to classes 1 and 2. Nearly all format conversions performed at seismological data centers are done to move upwards in the hierarchy for the purpose of data archiving and exchange with other data centers. Software tools are widely available to convert from one format to another and particularly upwards in the hierarchy.

This hierarchy also explains why there are so many formats. The design of class 1 formats depends on the manufacturer of the data acquisition system. In the early days of digital seismometry, display and analysis software was often proprietary and marketed specifically for a certain manufacturer's equipment and data format. There was no real need for manufacturers to adhere to a standard recording format, until users began to realize the advantages of exchanging data with other seismologists and discovered that this was quite difficult unless the other party was using the same hardware and/or software.

Station operators who were not satisfied with the proprietary analysis software supplied with the procured data acquisition systems started to convert data from Class 1 formats into the Class 2 formats which were used by more powerful and widely available analysis packages such as SAC. These programs usually provide subroutines that make conversion from local formats fairly easy. Analysis packages (e.g., SeisGram) which are developed around a Class 1 format (BDSN in this case) implicitly offer their format preference as a candidate for a new standard in Class 2, but it hardly matters as long as the necessary software tools are available to convert to and from the data exchange formats.

The GDSN (Global Digital Seismic Network) format began as a Class 1 format, but because it was used by an important global seismograph network (DWWSSN, SRO) it became accepted as a de facto standard for data exchange (Class 3). The beginning of widespread international data exchange within the FDSN (Federation of Digital Seismic networks) and GSE (Group of Scientific Experts) groups in the late 1980’s revealed the GDSN format's weaknesses in this role and put in motion the process of defining more capable exchange formats.

The volume of commonly available digital seismic data continues to increase dramatically. It increased from 600 MB annually in 1980 to 300 GB in 1992 and today we are talking about many terabytes. Database systems, which are specially designed to handle these large datasets, have therefore begun to appear as a superset of the standard data exchange formats. The SUDS system was an example of this type of format, however today it is little used.

In the 1990’s, several activities (e.g., the GSETT-3 experiment and the U.S. National Seismograph Network (USNSN)) emerged featuring real-time exchange of seismological data, and interest focused on formats which are suitable for such applications. In the late 1990’s, this idea was carried farther by systems such as Earthworm, which implemented format-independent protocols. Earthworm also is designed to exchange data across a peer network of multiple, independent nodes, as well as in a traditional network of dependent nodes with a centralized collection and distribution center.

Following is a brief description of some of the classes of formats as defined above.

10.3.1 Data archival

Data archival requires the storage of complete information on station, channel(s) and the structure of the data. Most existing formats are designed to provide part of the information. Most archival formats presently in use do include info on station and channel, but are not always complete in the description of the data. What we envisage is demonstrated through several features in the Standard for the Exchange of Earthquake Data (SEED) format:

- Data Description Language (DDL)
- Reference to byte order
- Response information

The DDL is defined to enable the data itself to be stored in any data format (integer, binary, compressed). The language consists of a number of keys defining, for example, the applied compression scheme, number of bytes per sample, mantissa and gain length in bits and the use of the sign convention. The reader interprets the DDL and knows exactly how to deal with the data. The advantage of the DDL is that the original data structure can be maintained and is known. A disadvantage is that readers will have to interpret the DDL and have less performance in reading. However, the decoding information is available directly with the data and this is extremely important, since data are collected on platforms having different byte orders. In SEED the byte order of the original data is defined in the header, so the reader will be able to decide whether bytes should be swapped.

In most archival formats, response information can be supplied in terms of poles and zeroes. Fewer efforts are undertaken to give the FIR filter coefficients in the header, although they are accounted for in the definition of SEED and GSE2.X. A problem occurs when a description of the instrument response is given only in measured amplitude and phase data as a function of frequency, as is the case in the GSE1.0 format. Also, the GSE2.X does not specify the minimum requirement. The main purpose of the response information is to correct for instrument response and thus the user will have to find the best fitting poles and zeroes to the given response. Although tools are available to calculate poles and zeroes from frequency, amplitude and phase data (e.g., in Preproc), results from the multiple inversion of the discrete frequency, amplitude and phase data will be different from the original data.

The deployment of large mobile arrays consisting of heterogeneous instrumentation is an important research tool. Data archival of these data is important. Although there is a tendency to store the data in a common format, the responses of sensors and data acquisition systems are often poorly known. It is recommended to pay attention to this issue before the experiment starts!

Finally, an issue in data archival is the responsibility of the data quality and the mechanism of reporting data errors. The network/station operator is responsible for the quality of the original data. However, the data may be subjected to format conversion at a remote data center. This last stage could introduce errors and it is the originator of the data, which must be responsible for data quality and should agree on the final conversion, if such a conversion is done externally.

10.3.2 Data exchange formats

The data exchange formats are closely related to the way data is exchanged. Therefore, these formats are described separately in this section. Essentially, any format can be used for exchange, however the idea of an exchange format is~~is~~ to make it easy to send it electronically, to have a minimum standard of content and be readable on all computer platforms. Today, the de-facto standard within the FDSN is the mini-SEED or SEED format (http://www.iris.edu/manuals/SEED_chpt1.htm).

At present, there are many different techniques in use to exchange data, either between data users and data centers or between data centers. Since the beginning of the 21st century techniques to transmit and exchange data in real-time became very popular and efficient, and the volume of continuous waveform data has grown very fast. With this rapid increase of continuous data, maintenance and adaptation of existing techniques and development of new techniques for requesting selected datasets only. Nowadays, the most common format which is used in real-time and request mechanisms for data exchange is (mini-)SEED. An overview of existing data request techniques is given below.

	Technique	**Advantage**	**Disadvantage**
Indirect on-line	autoDRM, NetDC, breq-fast	email based (no connection time)	small volume or download through ftp
Direct on-line	ftp, WWW, DRM (Wilber/FARM); DHI, ArcLink, webservices, QuakeExplorer	direct access, enables easy data selection	slow for large data volumes
Off-line	CD-ROM (DVD)	direct access	no real-time data

Indirect on-line data exchange is arranged through (automated) *Data Request Managers* (DRMs) where the request mechanism is based on email traffic. One of the earliest systems was AutoDRM (http://seismo.ethz.ch/autodrm), a standardized protocol using the GSE defined syntax. One step further is the implementation of a communication protocol for exchange between data centers in such a way that a user only has to send one request to a nearby data center node. His/her request is then automatically routed through the data centers that may contribute to the requested data set. Such a protocol is the NetDC system (Casey and Ahern, 1996). A similar system, called ArcLink, was developed by GFZ (GEOFON) using a more simple communication protocol over a TCP/IP socket, making it a direct on-line request mechanism

One basic problem in using email as the transport mechanism is the restricted data volume that can be exchanged. Also, the format sometimes will have to be ASCII. The format issue is taken care of in the GSE format, although in the description of the AutoDRM protocol it is mentioned that also a format like SEED can be used. The only difference is that the user is requested to get the data through anonymous ftp (pull) or the data is pushed into an anonymous ftp area defined by the user. The AutoDRM system at the Orfeus Data Centre

(ODC) supports both the SEED and GSE format in data exchange. Both NetDC and ArcLink provide mini-SEED waveform data.

Direct on-line access to data is arranged at the ODC (http://www.orfeus-eu.org/) for example through ftp, a web-interface and web services. Web services provide machine-accessible data services typically accessed through the standard HTTP protocol. In many cases these services are publicly exposed so that users may access them directly through their own applications. Other data centers, like IRIS-DMC and GEOFON, have similar web-based techniques. An example of an application using web services is the NERIES portal (http://www.neries-eu.org) through which a user can search for earthquake data in the EMSC database and download waveform data through the ODC.

Internet speed presently may still be limiting the usefulness of this direct on-line data exchange, especially since the volumes that are to be transferred may be huge. One major advantage of direct on-line availability of the data is the capability to make a selection out of the vast amount of digital data. Procedures are presently under development to increase the power of these selection tools, for example through the above mentioned portal and webservices developments.

Off-line data access provides complete, quality controlled data that are locally available at each institute in the form of CD-ROMs or DVD's. The completeness and quality control takes time and CD-ROMs and DVD's have a limited storage capacity.

10.3.3 Formats for data base systems

Formats for data base systems are specially designed and no details will be given here. Examples of such formats are CSS and the derived "IDC Database Schema" (see Information Sheet IS 10.3) and SUDS.

10.3.4 Continuous data protocols and formats

With better communication systems, real time transmission of digital data has become common. There is no internationally agreed upon protocol for this and equipment manufacturers often use their own formats. However, the number of protocols is limited, and the most common systems are SeedLink, SCREAM, NAQS, CD1.x, Antelope, EarthWorm, NRTS and LISS. A number of open-source and commercial data exchange systems exist between the different systems, and may be included in the protocol software or may be found on the appropriate web sites. Complete documentation for the CD-1.0 protocol used for the transmission of IMS data as described under 10.2.4. is now being replaced by the CD-1.1 protocol. Descriptions for both can be found on the secure web site https://www2.ctbto.org (authorized users only) and openly on http://www.geoinstr.com/pub/manuals/343r02p.pdf .

At present, a large number of channels from a station or array of stations can be transmitted in near-real time using a single connection. Digital data are provided in compressed or uncompressed format and with or without authentication signatures. The protocol uses units of information called frames to establish or alter a connection and to exchange data between the sender and the receiver. Only one frame is being transmitted or received at any instance. A time-out is used in case of lost connection.

Establishing connections. The sender initiates the connection with the receiver to a pre designated IP address and port by sending a Connection Request Frame. The receiver validates the authenticity of the sender and provides a new port and Internet Protocol (IP) address in a Port Assignment Frame. The sender drops the original connection and connects to the assigned IP address and port that is subsequently used for all data transfer.

Transmitting data. After the connection is established, the sender sends a Data Format Frame, which describes the format of the subsequent Data Frames. The sender can then send Data Frames data. The Data Format Frame provides information about itself and about Data Frames that will follow. The Data Frame contains the raw time series data. Each Data Frame has a single Data Frame Header and multiple channel sub-frames.

Altering connections. Either sender or the receiver can alter the connection through the exchange of Alert Frames. The receiver sends the Alert Frame to notify the sender to use a different port. The sender uses Alert Frames to notify the receiver that the communication will cease or that a new data format is about to be used.

Terminating connections. Typically, an established connection remains active and in use until the sender or receiver terminates it for maintenance or reconfiguration. The connection can be intentionally terminated by sending an Alert Frame. Unintentional termination due to a slow or failed communications system is detected after the time-out period.

Documentation for the Earthworm protocol can be found on the USGS website
http://folkworm.ceri.memphis.edu/ew-doc/ .

10.4 Some commonly encountered digital data formats

This section gives an alphabetical list of common formats in use by several recording or analysis programs. For each format some description is given. For several more specific, proprietary, now outdated or rarely used and platform formats we refer to Chapter 10, pages 12-16, in the first edition of NMSOP (DOI: 10.2312/GFZ.NMSOP_r1_ch10; http://nmsop.gfz-potsdam.de and http://www.isc.ac.uk/standards).

AH
Class: 2
The Ad Hoc (AH) format is used in the AH waveform analysis software package developed at Lamont Doherty Geological Observatory, New York, USA. This package also supports a number of conversion tools.

CSS
Class: 2,4
The Center for Seismic Studies (CSS) Database Management System (DBMS) was designed to facilitate storage and retrieval of seismic data for seismic monitoring of test ban treaties [CSS]. The seismic data separate into two categories: Waveform data and parametric data.

For the parameter data, the design utilizes a commercial relational database management system. Information is stored in relations that resemble flat, two-dimensional tables as in the ISF format (see Information Sheet IS 10.1). The description of waveform data is physically separated from the waveform data itself. The index to the waveform archive is maintained within the relational database. Data are stored in plain files, called non-DBMS files. Each non-DBMS file is indexed by a relation that contains information describing the data and the physical location of the data in the file system. Each waveform segment contains digital samples from only one station and one channel. The time of the first sample, the number of samples and the sample rate of the segment are noted in an index record. The index also defines in which file and where in the file the segment begins, and it identifies the station and channel names. A calibration value at a specified frequency is noted. The index records are maintained in the **wfdisc** relation. Each **wfdisc** record describes a specific waveform segment and contains an id number to designate detailed information on the station and instrumentation of the trace.

GSE

Class 3

The format proposed by the Group of Scientific Experts (GSE format) has been extensively used with the GSETT projects on disarmament. The GSE2.1, now renamed IMS1.0, is the most recent version. The manual can be downloaded from http://www.seismo.ethz.ch/prod/autodrm/manual/provisional_GSE2.1.pdf .

A GSE2.1 waveform data file consists of a waveform identification line (WID2) followed by the station line (STA2), the waveform information itself (DAT2), and a checksum of the data (CHK2) for each DAT2 section (Provisional GSE 2.1 Message Formats & Protocols, 1997). The default line length is 132 bytes. No line may be longer than 1024 bytes. The response data type allows the complete response to be given as a series of response groups than can be cascaded. Response description is made up of the CAL2 identification line plus one or more of the PAZ2, FAP2, GEN2, DIG2 and FIR2 response sections in any order.

Waveform identification line WID2 gives the date and time of the first data sample; the station, channel and auxiliary codes; the sub-format of the data, the number of samples and sample rate; the calibration of the instrument represented as the number of nanometers per digital count at the calibration period; the type of the instrument, and the horizontal and vertical orientation.

Line STA2 contains the network identifier, latitude and longitude of the station, reference coordinate system, elevation and emplacement depth.

Data section after DAT2 may be in any of six different sub-formats recognized in the GSE2.1 waveform format: INT, CM6, CM8, AUT, AU6, and AU8. INT is a simple ASCII sub-format, "CM" sub-formats are for compressed data and "AU" sub-formats are for authentication data. All represent the numbers as integers and therefore can be sent by email.

A checksum CHK2 must be provided in the GSE2.1 format. The checksum is computed from integer data values prior to converting them to any of the sub-formats.

SAC
Class 2
Seismic Analysis Code (SAC) is a general-purpose interactive program designed for the study of time sequential signals [SAC]. Emphasis has been placed on analysis tools used by research seismologists. A SAC data file contains a single data component recorded at a single seismic station. Each data file also contains a header record that describes the contents of that file. Certain header entries must be present (e.g., the number of data points, the file type, etc.). Others are always present for certain file types (e.g., sampling interval, start time, etc. for evenly spaced time series). Other header variables are simply informational and are not used directly by the program. Although the SAC analysis software only runs on Unix platforms and the general format is binary, there is also an ASCII version that can be used on any platform.

SUDS
Class: 1,2,4 Platform: PC
SUDS stands for "The Seismic Unified Data System". The SUDS format was launched to be a more well thought out format useful for both recording and analysis and independent of any particular equipment manufacturer. The format has seen widespread use, but has lost some momentum, partly because is a not made platform independent.

10.5 Format conversions

10.5.1 Why convert?

Ideally, we should all use the same format. Unfortunately, as the previous descriptions have shown, there are a large number for formats in use. With respect to parameter formats, one can get a long way with HYPO71, Nordic and GSE/ISF formats for which converters are available, such as in the SEISAN system. For waveform formats, the situation is more complicated. First of all, there are many different formats, and, since most are binary, there is the added complication that some will work on some computer platforms and not on others. This is particularly a problem with binary files containing real numbers as for example, the SeisGram format. Additional problems are: Some formats have seen slight changes and exist in different versions, different formats have different contents so not all parameters can be transferred from one format to another and conversion programs might not be fully tested for different combinations of data.

Many processing systems require a higher level format than the often primitive recording formats so that is probably the most common reason for conversion, and a similar reason is to move from one processing system to another.

The SEED format has become a success for archival and data exchange. Initially, it was not very useful for processing purposes, however now it is more widely used also for processing. Sometimes it is important to be able to move down in the hierarchy to be able to use a particular processing system. Thus, the main reasons for format conversion can be summarized as:

- Moving upwards in the hierarchy of formats for the purpose of data archiving and exchange

- Moving downward from the archive and exchange formats for analysis purposes
- Moving across the hierarchy for analysis purposes
- Moving from one computer platform to another

10.5.2 Ways to convert

There are essentially two ways of converting. The first is to request data from a data center in a particular format or to log into a data center and use one of their conversion programs. The other more common way is to use a conversion program on the local computer. Such conversion programs are available both as free standing software and as part of processing systems. Equipment manufactures will often supply at least a program to convert recorder data to some ASCII format and very often also to the more standard format MiniSeed.

10.5.3 Conversion programs

Since conversion programs are often related to analysis programs, we list in Tab. 10.6 some of the better-known analysis systems and the format they use directly.

Tab.10.6 Examples of popular analysis programs.

Program	Author(s)	Input format(s)	Output format(s)
Geotool	J.Coyne	CSS, SAC, GSE	CSS, SAC, GSE
SAC	IRIS	SAC	SAC
SEISAN	J.Havskov, L. Ottemöller, P.Voss	SEISAN, GSE	SEISAN, GSE, SAC
SeismicHandler	K.Stammler	miniSEED, GSE, AH, ESSTF, GCF	GSE, miniSEED
SNAP	M.Baer	SED, GSE	SED, GSE

An overview of available format conversion programs can be found on the ORFEUS Web pages under ORFEUS Seismological Software Library (http://orfeus.knmi.nl/ wirjung.groups/wg4/index.html). Here we present just a few packages in alphabetical order. Only those programs are mentioned which are able to read at least one of the formats mentioned in sub-chapter 10.4.

Codeco

Program **codeco** was written by U. Kradolfer and modified by K. Stammler and K. Koch. Input files can be in SAC binary or ASCII, or GSE formats. Output formats are: integer or compressed GSE1.0 or GSE2.0, SAC binary or ASCII, and miniSEED. **Codeco** is available through the SZGRF software library (ftp://ftp.szgrf.bgr.de/pub/software).

GSE to SEED

Program **gse2seed**, developed by R. Sleeman (Orfeus Data Centre, de Bilt), converts a GSE2.X file to the SEED2.3 format. Multiple traces are handled. For each WID2 section, the GSE file must contain corresponding data types STATION, CHANNEL and RESPONSE. This program was originally developed to convert both metadata and waveform data, but is maintained nowadays only to convert GSE metadata into dataless SEED. Conversion of GSE waveform data into mini-SEED can be done by **gse2mseed**, developed by Chad Trabant (IRIS DMC).

PASSCAL package

The PASSCAL package was written by P. Friberg, S. Hellman, and J.Webber, developed on SUN under SunOs4.1.4, compiled under Solaris 2.4 and higher and also under LINUX. It converts RefTek to SEGY and miniSEED. Program **pql** provides a quick and easy way to view SEGY, SAC, miniSEED or AH seismic data. **pql** operates in the X11 window environment. The package is available from the PASSCAL instrument center (http://www.passcal.nmt.edu) at New Mexico Tech., Socorro.

Preproc

Preproc has been designed to assist the seismologist who wishes to analyze large sets of raw digital data that need to be preprocessed in some standard way prior to the analysis. Preproc was written by Miroslav Zmeskal for the ISOP project in the period 1991-1993. It was rewritten recently in the object-oriented form. As a by-product, **preproc** can perform data conversion from GSE / PITSA ISAM to GSE / PITSA ISAM. In the near future new input/output formats will be implemented (ESSTF, miniSEED). **preproc** was successfully compiled on HP, SUN, Linux and DOS. Program package **preproc** and a detailed manual are available through the ORFEUS Seismological Software Library

Rdseed

Rdseed reads from the input tape or file in the SEED format. According to the command line function option specified by the user, **rdseed** will read the volume and recover the volume table of contents (-c), the set of abbreviation dictionaries (-a), or station and channel information and instrument response table (-s). In order to extract data from the SEED volume for analysis by other packages, the user must run **rdseed** in user prompt mode (without any command line options). As data is extracted from the SEED volume, **rdseed** looks at the orientation and sensitivity of each channel and corrects the header information on request. Implemented output formats are (option d): SAC, AH, CSS 3.0, miniSEED and SEED. A Java version of rdseed is to be released in 2001. **Rdseed** was developed by Dennis O'Neill and Allen Nance, IRIS DMC.

SeedStuff is a set of basic programs provided by the GEOFON DMS software library in Potsdam (ftp://ftp.gfz-potsdam.de/pub/home/st/GEOFON/software) to process and compile raw data from Quanterra, Comserv and RefTek data loggers. The goal is to check and extract data from station files/tapes to miniSEED files and to assemble miniSEED files to full SEED volumes. The SeedStuff package was written by Winfried Hanka and compiled on the SUN, HP and Linux. The following tools are available:

extr_qic: extracts multiplexed raw Quanterra station tapes to demultiplexed miniSEED files containing only one station / stream / component
extr_file: like extr_qic for multiplexed miniSEED, RefTec files
extr_fseed: disassemble full SEED tapes. SEED headers are skipped, data are stored into station / stream / component files
check_seed: checks the contents of miniSEED data files or tapes
check_qic: analysis the contents of a Quanterra data tape
copy_seed: assembles a full SEED volumes from miniSEED files for a given set of station / stream / component defined in the copy_seed.cfg configuration file
make_dlsv: generates a dataless (header only) SEED volume for a set of station/stream/component defined in copy_seed.cfg

SEED to GSE

Recently (2009), a SEED to GSE converter was developed at ORFEUS Data Center (ODC), mainly to support the GSE format within AutoDRM at ODC. This tool will become available at the ORFEUS software library.

SEISAN

The SEISAN analysis system has about 40 conversion programs, mostly from some binary format to SEISAN. The SEISAN format can then be converted to any standard format like SEED, SAC or GSE. SEISAN has format converters for most recorders on the market including Kinemetrics, Nanometrics, Teledyne, GeoSig, Reftek, Lennartz, Güralp and Sprengnether.

Acknowledgement

The authors acknowledge with thanks the careful review by Bruce Presgrave of the US Geological Survey. It has improved both the language of the original draft and provided useful references to the Earthworm system. In addition we acknowledge discussion and comments from Lars Ottemöller and Reinoud Sleeman.

Special references

- [CSS] Anderson, J., W. Farrell, K. Garcia, J. Given, and H. Swanger, Center for Seismic Studies Version 3 Database: Schema Reference Manual, SAIC Technical Report C90-01, 1990.
- [IDC3.4.1] Formats and Protocols for Messages, Rev. 3, 2001.
- [GSE] Provisional GSE2.1 Message Formats & Protocols, 1997. Operations Annex 3, GSETT-3.

- [LEN] SAS-58000 User's Guide and Reference Manual, 1986. Lennartz electronic GmBH
- [SAC] W.C. Tapley & J.E. Tull, 1992. SAC - Seismic Analysis Code. LLNL, Regents of the University of California
- [SEED] Standard for the Exchange of Earthquake Data, 1992. Reference Manual, SEEDFormat v2.3, FDSN, IRIS, USGS

References

Casey, R., and Ahern, T. (1996). Technical manual for Networked Data Centers (NETDC) protocol (*IRIS, internal report*) or http://www.iris.washington.edu/manuals/netdc/

IASPEI (2005 and 2011). Summary of Magnitude Working Group recommendations on standard procedures for determining earthquake magnitudes from digital data. Preliminary version October 2005, updated version September 2011; http://www.iaspei.org/commissions/CSOI.html.

Klein, F. W. (1978). Hypocenter location program HYPOINVERSE. *U.S. Geol. Surv. Open-File Report.* **78-694**.

Lahr, J. C. (1981) REFERENCE IS MISSING!

Lee, W. H. K., and Lahr, J. C. (1975). HYPO71 (revised): A computer program for determining hypocenter, magnitude and first motion pattern of local earthquakes. U.S. Geological Survey Open-File Report 75-311, 116 pp.

Willmore, P. L. (Ed.) (1979). Manual of Seismological Observatory Practice. *World Data Center A for Solid Earth Geophysics,* Report **SE-20**, September 1979, Boulder, Colorado, 165 pp.