LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY
- LIGO -
CALIFORNIA INSTITUTE OF TECHNOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Specification	**LIGO-T980070-02 - E**	**07-29-98**
LIGO Metadata, Event and Reduced Data Requirements PRELIMINARY		
LIGO Data Analysis Group		

This is an internal working note

of the LIGO Project

California Institute of Technology
LIGO Project - MS 51-33
Pasadena CA 91125
Phone 1.626.395.2129
Fax 1.626.304.9834
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Project - MS 20B-145
Cambridge, MA 01239
Phone 1.617.253.4824
Fax 1.617.253.7014
E-mail: info@ligo.mit.edu

WWW: http://www.ligo.caltech.edu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE AND ACRONYMS

1 Scope

This document defines LIGO metadata, reduced data and event data. It document complements other LIGO documents specifying LIGO Frame Data and Lightweight Data formats.

2 Applicable Documents

The documents cited in ***Table 1: Relevant Documents*** serve as reference material.

Table 1: Relevant Documents

DOCUMENT TITLE	DATE AND ID NUMBER
A Light-Weight Data Format for LIGO	Version 1.1 LIGO-TBD
Global Diagnostics System Preliminary Design	T970172
LIGO Frame Format Specification	T970130
LDAS Design Requirements Document	T970159
LDAS Conceptual Design Document	T970160

3 LIGO Metadata

3.1. Definition

Metadata are data about data. They are data not immediately derivable from a particular dataset itself; contextual information relevant to interpreting that dataset. LIGO Metadata includes, but is not limited to, the following elements:

Tape or other media ID numbers;

- Time/date tags when particular data were acquired or generated;
- Originator or creator of the data;
- Location of the data (URL or file system identification);
- Volume of data; format of data;
- Location or identification of other relevant or related data (e.g., calibrations, trends, statistical descriptors, etc.);
- Relationship of a dataset to another dataset;
- Descriptive textual information about a dataset;
- etc.

LIGO metadata will be stored in XML according to the format specified at URL http://www.cacr.caltech.edu/~roy/ligo/xml.

3.2. LIGO Raw Data Catalog

The principal archive of *raw* LIGO data will be in the form of data frames. Data are written to frames as part of the acquisition process at the observatories. For LIGO I operations, it is expected that data will transferred from on-line mass storage (spinning media) to physical, permanent media (e.g., tapes, dvd, etc.). The media will be transferred from the observatories to the long term archival center (CACR) on a regular basis using commercial carrier means of transportation. At some future date, it may become feasible to transmit data electronically; however this is not presently the baseline.

The raw data archive is necessarily divided into three components (data ages are comparative):

- Data at the observatories.
  - *Current data* [T < 16 hours]: These are the on-line data and their associated metadata which are relatively recent (T < 1 day old) are still on the local on-line disk system.
  - *Recent data* [16 hours < T < 30 days]: These are data on (non-spinning) media and their associated metadata which were recently removed from the on-line mass storage system. These will be copies of data being shipped to the archive. The media shall be preserved at the observatories at least until the same data have been ingested into the long term archive. A means shall be provided to perform a "handshake" on a regular basis to identify media at the observatories which may be recycled or otherwise reused.
- Data in transit. Data will be in transit from observatories to the archive for one to several days. These data will be *recent data* (see above) and will only be available from the observatories.
- Data at the archive.
  - *Data in process* [T < 30 days]: Data just received from the observatories awaiting processing. The associated metadata remain at the observatories. Raw data will also only be generally available from observatories.
  - *Archived data* [T > 30 days]: Once raw (frame) data have been reduced and/or ingested into the archive, the associated metadata will be transferred from the observatories to the archival site. This will be done on a regular basis using electronic transmission. At this point redundant data at observatories may be purged.

The primary form for raw data is in frame format. Frame data are grouped into entities termed FrameDataSets (FDS). A FDS is an object consisting of a number of frames taken *contiguously* during a given epoch. A FDS is an object for which all frames contained within it have the following common characteristics:

- The same set of channels with corresponding channel attributes (e.g. name, gain, etc.);
- The same "primary" instrument state (the instrument state vector will be denoted as having "primary" and "secondary" elements);
- The same interferometer(s);
- Contiguity -- one FDS may not contain or span other FDSs. [This needs clarification -- does loss of lock necessarily force an FDS to end and a new one to begin?];
- One media may contain more than one FDS;
- One FDS may span multiple media;

- Metadata about an FDS will include:
  - FrameDataSet ID
  - FDS start time, stop time and duration (TStart, TStop, DeltaT);
  - Media ID on which FDS is located/spans;
  - Frame count and frame type in FDS;
  - URL or other source ID for relevant (collateral) data service:
    - Operator or other experimental logs and comments;
    - Calibration files (typically also be contained within frame data);
    - Trigger data, veto data or other diagnostic data acquired during the same epoch
    - Non-LIGO data (seismic data; weather data; etc.)
  - Server ID for any event or reduced data generated by previous analyses of the same FDS.
  - IDs of other (LSC) users who have "touched" the same FDS.

3.3. LIGO Event Data Catalog

Results of particular data analyses will result in "events". The vast majority, if not all, of these events will have instrumental, geophysical, or anthropomorphic antecedents. These antecedents are expected to have been picked up as triggers or vetoes by other, non-interferometric, sensing channels. The events, taken in the context of the analyses which generated them and the ancillary data which may explain them, will be archived as a LIGO Event Data Catalog. Eventually, astrophysically significant events will also become included in this archive.

It is possible to write events as frames; however it is likely that the need for easy and universal access will be such that events will become objects either incorporated in or indexed by a database. The first events will be those detected by on-line algorithms at the observatories. These will include diagnostics-based triggers or vetoes. In addition, on-line searches for astrophysical signals will also generate events. Event rates will be low enough so that they may be archived essentially indefinitely at the observatory sites. Moreover, they can be transmitted over the LIGO WAN to the LIGO Archive at Caltech. Later events will be generated by off-line analysis and processing.

The event data archive is divided into at least four components:

- Event data at the observatories. There will be on-line event data which are specific to the observatory acquiring the data. Metadata for these event data shall reside at the respective observatories initially. Thus, part of the event metadatabase will need to be distributed across several LIGO sites.
- Event data at the archive. Off-line analyses will also generate event data. These will be archived at the LIGO Archive at Caltech. It is also likely that the EventDataSets (EDSs) acquired at the observatories will be transferred to a common repository so that events from all LIGO interferometers can be brought together in a unified sense.
- Event data at LIGO Laboratory sites. Off-line data analysis at LIGO Laboratory sites (observatories, Caltech, MIT) will also generate events. "Significant" or "useful" events (TBD) will be available as part of the event data catalog. It may also be the case that such events may be transmitted to the archive.
- Event data at collaborating institutions. Members of the LSC may also be performing data analysis remotely. Events generated by these analyses will be treated in a manner similar to

events generated at LIGO Laboratory sites,

Events from individual LIGO interferometers which occur within the coincidence window for the LIGO Detector will be grouped into an entity termed an Event. An Event is an object consisting of all relevant data for the epoch during which it was detected.

Metadata about the Events will include:

- Event format (e.g., frame or lightweight format)
- FDS IDs (for multiple interferometers) from which Event was generated;
- Algorithm or other parameters used to generate the event;
- Importance of event; whether the event has been explained -- pointer or link to metadata providing explanation;
- Event start time, stop time and duration (TStart, TStop, DeltaT), as appropriate for the event:
- Media ID on which the corresponding FDS is located/spans;
- URL or other source ID for relevant (collateral) data:
  - How was event generated? By whom? LSC user ID(s) of event generator/analyzer;
  - Location of raw data “snippet” containing the event in question;
  - Location of operator or other experimental logs and comments;
  - Location of calibration files;
  - Location of trigger data, veto data, and other diagnostic data acquired during the same epoch
  - Location of non-LIGO data (seismic data; weather data; etc.)

3.4. LIGO Reduced Data Sets

Reduced data are derived from the raw data by any number of (linear and non-linear) filtering techniques:

- RMS filtering;
- Maximum/minimum filtering;
- Mean/median filtering;
- Gating on state vector condition (e.g., locked/unlocked state);
- Decimation and trending over long periods of time;
- etc.

Certain reduced data will be generated as part of the acquisition process and written as trend frames. Additional reduced or trend data will be generated during the ingestion process as the latest data are incorporated into the archive.

Some reduced data will be have sufficient utility to be stored separately from the raw data (e. g., as trend data). This will be so either because the algorithms used to create them may be complex and not easily or often repeated, or possibly because the reduced data are particularly useful representations of the data and they may themselves be efficiently used for subsequent analysis and filtering.

Reduced data are grouped into an entity termed a ReducedDataSet (RDS) whose functionality is intended to be analogous to the FDS. Typically a RDS consists of frame data. A RDS is an object consisting of a number of *contiguous* frames taken during a given epoch. A RDS is an object for which all frames contained within it have the following common characteristics:

- The same set of channels with corresponding channel attributes (e.g. name, gain, etc.);
- Contiguity -- one RDS may not contain or span other RDSs;
- One media may contain more than one RDS;
- One RDS may span multiple media;
- Metadata about an RDS will include:
  - ReducedDataSet ID
  - RDS start time, stop time and duration (TStart, TStop, DeltaT);
  - Media ID on which RDS is located/spans;
  - Frame count and frame type in RDS;
  - URL or other source ID for relevant (collateral) data:
    - Location of operator or other experimental logs and comments;
    - Location of calibration files;
    - Location of trigger data, veto data, and other diagnostic data acquired during the same epoch
    - Location of non-LIGO data (seismic data; weather data; etc.)
  - Location and ID of any event data generated by previous analyses of the RDS.
  - LSC user IDs of others who have “touched” the same RDS.

Metadata about the reduced data will include:

- RDS ID from which reduced data were generated;
- Algorithm or other parameters used to generate the reduced data;
- Time tag (TStart, TStop, DeltaT), as appropriate for the reduced data:
- Media ID on which the corresponding RDS is located/spans;
- URL or other source ID for relevant (collateral) data:
  - Location of operator or other experimental logs and comments;
  - Location of calibration files;

4 Indexing methods into the LIGO Data Archive

4.1. Access by time

Using GPS or UTC, retrieve an FDS, Event or RDS or portions thereof. Indicate where/how data are to be returned, stored or handed off to an MPI process (batch mode). Retrieval is specified by:

- data volume:
  - {start time, stop time} , or
  - {start time, Δtime}, or

  - {start time; number of frames};
- transfer destination:
  - path, URL, socket, or filtering process;
  - object or file name for data transfer;
  - data format for transfer;

4.2. Access by detector orientation

Using an ephemeris database, retrieve an FDS or RDS or portions thereof according to how the detector array (LIGO or LIGO-plus-others) was aligned relative to celestial (or galactic) coordinates. Indicate where/how data are to be returned, stored or handed off to an MPI process (batch mode).

Retrieval may be specified by:

- orientation:
  - angular coordinates of antenna pattern peak senstivity;
- frequency shift rates for periodic sources in specified positions on the sky;
- destination for data transfer:
  - path, URL, socket, or filtering process;
  - object or file name for data transfer;
  - data format for transfer;

4.3. Access by trend or reduced data characteristics

Using reduced or trend data when these are available, provide “cuts” on FDSs, EDSs or RDSs according to their reduced data properties. An example of this would be to return all FDSs spanning a given epoch for which the RMS in a certain channel (or channels) is (are) within specified bounds. Other forms of returned results may include: statistical descriptors of bulk data -- RMS-vs-time; RMS-vs-frequency of occurence; correlations between trend data from different channels; etc.

4.4. Access by trigger, veto or other event data

Using event data (and metadata), access all FDSs or RDSs spanning the epoch during which certain criteria are met on triggers, vetoes or other event data.