LIGO Laboratory / LIGO Scientific Collaboration

LIGO-M1000066-v25 LIGO Laboratory June 2017

LIGO Data Management Plan, June 2017

http://dcc.ligo.org/LIGO-M1000066/public/

Stuart Anderson, Roy Williams

California Institute of Technology
LIGO Laboratory – MS 100-36
1200 E. California Blvd.
Pasadena, CA 91125
Phone 626-395-2129
Email: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Laboratory – NW22-295
185 Albany St
Cambridge, MA 02139
Phone 617-253-4824
Email: info@ligo.mit.edu

LIGO Hanford Observatory
P.O. Box 159
Richland WA 99352
Phone 509-372-8106

LIGO Livingston Observatory
P.O. Box 940
Livingston, LA 70754
Phone 225-686-3100

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

http://www.ligo.org

LIGO DATA MANAGEMENT PLAN

1 Introduction 3
1.1 The LIGO Scientific Collaboration 4
1.2 Open Data and Preservation 5
1.2.1 Phase 1: Detection or Discovery Phase *5*
1.2.2 Phase 2: Observational Phase *5*
1.3 Transition to Open Data 6
1.3.1 Preparation *6*
1.3.2 Public data release *6*
1.4 LIGO as an "Open Archival Information System" 8
1.5 Scope of Data Release 9
1.6 Unified Data Management Plans 9
2 Advanced LIGO 10
2.1 Observing Runs 10
2.2 Timeline 10
3 Work plan 11
3.1 Ingest 11
3.2 Storage 12
3.3 Metadata 12
3.3.1 Calibration *13*
3.3.2 Data Services *13*
3.4 Preservation 13
3.4.1 Preserving the Bits *13*
3.4.2 Preserving the Meaning *14*
3.4.3 Data Curation and Publication *14*
3.5 Operations 14
3.5.1 LIGO Data Grid *14*
3.5.2 Open Data Delivery *14*
3.5.3 Cloud Computing *14*
3.6 LSC Data Access and Computing 15
3.6.1 Data Access *15*
3.6.2 Software *15*
3.6.3 Help and Documentation *16*
3.6.4 Management *16*
3.7 Public Open Data Access 16
3.7.1 Data Services *17*
3.7.2 Data with Quality Information *17*
3.7.3 Interoperability *18*
3.7.4 Community *18*
3.7.5 Software *19*
4 References 19
5 List of Acronyms 20
Appendix A: update January 2012 21
5.1 A.1 Progress during 2011 21
A.1.1 User requirements much better defined *21*
A.1.2 LOSC proposal submitted *21*
A.1.3 Data publication with article *22*
5.2 A.2 Plans for 2012 22
A.2.1 Engineering runs now defined *22*

A.2.2 Trigger Delivery 22
A.2.3 Second Open Data Workshop 22
5.3 A.3 Changes to Data Management Plan 22
Appendix B: update January 2013 23
5.4 B.1 Progress during 2012 23
B.1.1 LOSC awarded and started 23
B.1.2 Data release for GRB051103 23
B.1.3 Website prototype 23
5.5 B.2 Plans for 2013 23
B.2.1 Building the LOSC archive 23
B.2.2 Engineering runs underway 23
B.2.3 Second Open Data Workshop 23
5.6 B.3 Changes to Data Management Plan 23
Appendix C: update February 2014 24
5.7 C.1 Progress during 2013 24
C.1.1 S5 open data is near release 24
C.1.2 Engineering Runs continue 24
C.1.3 Rapid alerts and observational follow-up of LIGO alerts 24
5.8 C.2 Plans for 2014 24
C.2.1 Review and release data from initial LIGO (S5 run, 2005-2007) 24
C.2.2 Prepare S6 for release pending feedback from the S5 release 24
C.2.3 Webinar tutorials about LOSC data products 24
C.2.4 Prototyping the Rapid Follow-up system in Engineering Runs 24
Appendix D: update January 2015 24
5.9 D.1 Progress during 2014 25
D.1.1 S5 open data is released 25
D.1.2 Future LIGO Runs 25
5.10 D.2 Plans for 2015 25
D.2.1 Review and release data from S6 25
D.2.2 Prototyping release of Advanced LIGO data 25
Appendix E: update August 2016 25
5.11 E.1 Progress during 2015 25
5.12 E.2 Plans for 2016 26
Appendix F: update May 2017 26
5.13 F.1 Progress since Aug 2016 26
5.14 F.2 Future Plans 26

1 Introduction

This Data Management Plan (DMP) is a deliverable as stipulated in the Cooperative Agreement with the National Science Foundation for the operation of LIGO. This plan describes how data from the LIGO instruments will be ingested, stored, made available and useful to the community of users of the data, and preserved for future use. In addition, LIGO is working with the Gravitational Wave International Committee (GWIC), and we have engaged our international partners, to coordinate broader access to all gravitational wave data. However, this plan addresses only the LIGO component. The plan is intended to be a living document, updated for the NSF on a regular basis, and this version (May 2017) is the seventh version. This latest version is publicly available at http://dcc.ligo.org/LIGO-M1000066/public/.

Appendix F contains a report on progress since Aug 2016 and updates to the plan for future activities.

The context for this plan is as follows: (1) Advanced LIGO is running at an astrophysically interesting sensitivity, but the evolution of the instrument performance in the future is difficult to deterministically predict; (2) LIGO has detected and published its first gravitational wave events (GW150914, GW151226, the candidate event LVT151012, and GW170104); (3) Observational and theoretical progress in the understanding of gravitational wave sources in the next few years may yield significant changes in our current views. Therefore, the assumptions underlying this plan will almost certainly evolve. Subsequent versions of this plan will reference these changes in knowledge as justifications for any modifications of specific aspects of the plan.

The plan discusses how we will make future LIGO data open to the broader research community. This plan calls for a phased approach to the release of LIGO data moving from the initial detection era to the Advanced LIGO epoch of routine detections. An outline of this plan has already been released and reviewed by the NSF; this paper is an elaboration of that initial plan [1] including modifications requested by NSF. We envision two broad phases of the scientific research and their associated data policies: the Detection Phase, and the Observational Phase.

We will outline how data has been available to the collaboration (LSC), and how we have transitioned to open data starting with the release of the Initial LIGO Science data in 2014 and 2015. We describe the Open Archival Information System (OAIS) [2][3] model of a data archive, what data are to be released, and discuss the powerful computing needed to extract the science from LIGO data. Section 2 outlines the broad process of work for the Data Management Team in the next years. Section 3 discusses how the six OAIS components relate to LIGO data management, and describes the "LIGO Open Science Center", that has been operational since 2012.

This plan is only for LIGO data. Collaborating gravitational-wave detectors, e.g., Virgo, and KAGRA, will maintain their own data management plans. LIGO will strive to make arrangements with these Collaborations that optimize the broad utility of the combined data set.

1.1 The LIGO Scientific Collaboration

To promote the best possible scientific use of LIGO data, the LIGO Scientific Collaboration (LSC) has worked for more than a decade to develop and refine data characterization techniques, data quality vetoes, and extensive analysis pipelines that take the raw data and produce astrophysical results. The LSC [4] is an *open* collaboration, allowing scientists from the global community to become involved in observations with LIGO data, based on their willingness to contribute to LIGO. Any scientist who is motivated to utilize LIGO data may do so by becoming a member of the LSC. The openness and success of this model is evidenced by the continuing growth of the collaboration, as well as its breadth. As of May 2017, it numbers 102 institutions worldwide with over 1150 members. The LSC has established a recognized international leadership in the field; and it has forged cross-collaboration international memoranda of understanding (MOUs) to establish a global network of comparably sensitive interferometric detectors, of which the LIGO instruments are key.

While data from Initial and Enhanced LIGO have been released openly, membership in the LSC is currently the only way to access Advanced LIGO data for analysis. Any astrophysical publications based on that data must first be approved by the collaboration and then, according to the collaboration's bylaws, will be published with a full list of authors.

1.2 Open Data and Preservation

1.2.1 Phase 1: Detection or Discovery Phase

Even with the initial discoveries of gravitational waves the understanding of instrumental artifacts and other spurious influences on the data is still evolving. While the instruments are evolving toward design sensitivity, data release to the broader research community makes the most sense for *validated* gravitational wave data surrounding confirmed discoveries, such as the release of data associated with GW150914 in February 2016, and the releases associated with GW151226 and LVT151012 in June 2016, and the release of GW170104 in June 2017. In addition, LIGO will release data corresponding to important non-detections where detectable gravitational waves might plausibly have been expected, e.g., the data surrounding one or more gamma-ray bursts, which are surprising in some way, or supplying evidence for or against a scientific model. In this phase, detected events are publicly released, in a manner analogous to astronomical post-processed image data, for scientific use by the broader research community. Release of events and important non-detections will occur with publication of one or more papers discussing these observational results in the scientific peer-reviewed literature. It should be noted that the first few LIGO detections were very carefully reviewed, and required several months of vetting; such care is justified for the first few detections. In addition to the four event data releases, two important non-detections have been published: a gamma-ray burst with coincident GW upper limit (GRB 051103), and the results of an Initial LIGO blind injection (the so-called 'big dog' event).

In addition to these non-detection data releases LIGO has released datasets described in the table below. The LIGO Open Science Center (LOSC) has provided the archive (OAIS implementation, see below), and the LSC has performed an internal data review to vet the released data.

Event and Trigger Open Data	Gigabytes	Release date
GW150914	4	2016 Q1
GW151226	4	2016 Q2
LVT151012	4	2016 Q2
GW170104	4	2017 Q2
Observation runs		
S5	7,700	2014 Q3
S6	2,300	2015 Q2

For full information see ref [21] or visit the website **http://losc.ligo.org**.

Detected events in this phase are released with sufficient data such that the broader community can comprehend the reported signals. There will be a minimum of 4096 seconds of strain data released, if available, so that open data users can adequately assess noise levels. Examples of such use might be analysis with a custom template bank by a researcher, finding correlations of LIGO data with other event streams, searching for afterglow, outreach, or an artist translating the data into music. For short duration events or very narrow band-limited continuous wave signals, calibrated time series containing the detections will be published along with associated metadata, which will include detected and reconstructed parameter estimates, with their uncertainties. In the case of a stochastic signal, a sky map will be released. In these cases, the quantity of data will be modest, and a detection portfolio should easily be downloadable to a modern laptop computer.

1.2.2 Phase 2: Observational Phase

The LIGO Laboratory and the LSC have worked closely with the broader research community to establish requirements, then build and field-test data access methods and tools. These methods and tools are provided to the broader community in Phase 2. As LIGO transitions past the first few detections, the understanding of the data has improved and the field is moving from the novelty of initial discoveries

toward the more mature exploration of the astrophysical content of gravitational wave signals. Use of population statistics, spatial distributions and the full power of multi-messenger astronomy will become the norm. During this epoch the entire body of gravitational wave data, with data-quality flags to show instrumental idiosyncrasies and environmental perturbations, will be released to the broader research community. In addition, LIGO will begin to release near-real-time alerts to any interested observatories as soon as LIGO "may" have detected a signal.

We note that the Phase 2 data release is part of the long-term data preservation plan for LIGO. By building the tools, data products, and documentation carefully, we ensure that this very valuable data will be available to future scientists, extracting further scientific value.

Even though Advanced LIGO's first and second observing runs (O1 and O2) are in Phase 1, it is planned to release their data at the end of each run as was done for Initial LIGO's S5 and S6 observations. In particular, O1 data are in the final stages of review and will be publicly released no later than two years after the end of O1 (January 2018). The O2 run is currently underway (see Ref [24]) and data will be publicly released no later than 18 months after the end of that run.

1.3 Transition to Open Data

The transition to Open Data, with the regular release of data during observation runs and prompt public alerts of transient events, will happen between the O2 and O3 runs. O2 is currently schedule to end in the third quarter of 2017, with O3 starting 1-2 years later.

1.3.1 Preparation

The LIGO Open Science Center was created to build data products suitable for public release (losc.ligo.org). LOSC has canvassed the community on requirements for open data (see Appendix A); we have built prototypes and basic tools, and documentation so that individual researchers can each comprehend the data according to their expertise. We are working from existing archive technology, for example the NASA archives Lambda [18], IRSA [19], and HEASARC [20] to design the LIGO archive. We are encouraging the LSC to become involved in the open data prototypes as an effective interface to LIGO data and algorithms.

In the years 2012-2016, we have built a sophisticated web application to return data either to a human making selections, or by web services with machine-readable output. Strain data around each event is in various formats, including HDF5, with included metadata and data quality flags. Source parameters of events are stored in machine-readable form, and presented in web pages, event catalogs, and a variety of dynamically generated plots.

In addition to releasing an hour or so of data around each event, the LOSC has released Initial LIGO data (S5 and S6 runs 2005 - 2010), and soon will release the data for the O1 run (Sep 2015 - Jan 2016). The strain data is converted to 4096-second duration files, with various formats, decimated from 16 kHz to 4 kHz sample rate, with the addition of data quality and injection information at 1 Hz.

1.3.2 Public data release

LIGO has decided to make the transition to an "open data observatory" after the end of the O2 run, which will be sometime in the third quarter of 2017 and before the start of O3 1-2 years later.

Once the transition to open data is made, regular public release of full time series data will occur. At this stage, LIGO will release all calibrated strain data, data quality flags, and times of hardware injections. Releases will occur every 6 months, in blocks of 6 months of data, with a latency of 18 months from the end of acquisition of each observing block. This 18-month proprietary data period between acquisition and release is required to validate, calibrate and process the data internally by the collaboration. It is expected that this proprietary period, which has already been reduced from an initial 24 months will be

further reduced in the future as improved understanding warrants. Even after LIGO enters Phase 2 we will continue releasing detection data at the time of publication, as described for Phase 1. In Phase 2, we will also start a program of public transient alerts (VOEvents, see sections 3.6.1.2, 3.7.1.2 and 3.7.3.2) so that observers around the world can attempt rapid follow-up and afterglow detections. This kind of program was extremely successful in elucidating the nature of gamma-ray bursts in the 1990's through the NASA GCN infrastructure [5]. This is an extension of the current practice, which is already underway with selected observatories via MOUs.

Releasing h[t] and related products requires a continuous and regular campaign of data quality assurance. Currently, this takes place within LSC working groups. There is no standard process for creating the annotated h[t] stream as a final product: it is performed within the software pipelines that have been developed. We note that the concept of a single "cleaned" h[t] stream is inaccurate: suitability for analysis h[t] depends on the particular search involved. Currently, during the course of a science run, generation of h[t] data is an iterative process wherein new versions of h[t] and the data quality flags are generated each time the calibration is refined. Within the LSC, it has been possible to deprecate old versions of h[t] with less accurate calibrations because the data release and access is controlled. Once we are in an era of open data releases, this will be more difficult; thus the time it takes to converge on a final calibration sets one time scale for the period before data are made public (see discussion below).

It should be noted that LIGO's open data policy is consistent with policies currently in place at NASA. Based on LIGO experience, the process of annotating the data (identifying artifacts, tagging hardware injections, providing accurate calibrations) and performing analyses takes up to 18 months. At this point, an eighteen-month delay before public release is required for vetted gravitational wave data, at least for the first data from Advanced LIGO. For comparison, NASA policy varies, but typically there is a 12-month delay in data releases, during which time mission scientists have access to the data for analysis, vetting, and clean up before they are made available to the broader community. However, it should be noted that, unlike LIGO, NASA has a history of doing public data releases. The first NASA mission to release data was the Cosmic Background Explorer (COBE) in 1990. The delay in release for COBE data was 36 months.

During 2014 and 2015, we released data from the initial LIGO S5 and S6 observational runs ahead of our plan, as described in detail in [21]. The LOSC team (authors of [21]) began work in 2012, focusing on an initial set of milestones for documenting, distributing, and curating the LIGO data:

• releasing the full GW strain data from the 2005–2007 "S5" science run, which achieved design sensitivity for Initial LIGO;

• releasing the full GW strain data from the 2009–2010 "S6" science run, which achieved design sensitivity for Enhanced LIGO;

• releasing the full GW strain data for the three published detections and one candidate event, annotated with data-quality flags;

• validating the GW strain data and annotating it with easily interpreted data-quality information and with information about the simulated GW signals present in the data (a.k.a. hardware injections);

• setting up the hardware and software for a LOSC web portal, losc.ligo.org, offering access to open data as well as additional services;

• providing a database service for the available data and its quality;

• producing a suite of tutorials and example codes to acquire, load, manipulate, filter, and plot the data.

It should be noted that the LOSC is jointly charged by both the LIGO Laboratory and LIGO Scientific Collaboration with its mission to provide open data to both the general public and LSC members. The S5

and S6 data has just one channel -- the strain channel -- downsampled and with bad data removed. It does not include the tens of thousands of channels describing the detectors and physical environment.

1.4 LIGO as an "Open Archival Information System"

Storage, access and preservation are intertwined aspects of data management, and LIGO expects to do all three, compliant with an international standard: Open Archival Information System, ISO 14721, [2][3]. An OAIS is an archive, consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community. There is a reference model and requirements, defined in ISO 14721 from 2003 [2]. In planning the LIGO data archive, our design fits with the requirements of OAIS, as in the subsections of section 3 that cover Ingest, Storage, Metadata, Preservation, Operations, and Access.

The OAIS model calls for definition of a "Designated User Community", who will be using the archive; the information to be preserved should be "independently understandable" to the designated community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information. The OAIS archive should also follow documented policies and procedures to ensure the information is preserved against all reasonable contingencies; remain accessible to the Designated User Community over the long term even if the original computing environment becomes obsolete; and to enable dissemination of authenticated copies of the preserved information in its original form, or in a form traceable to the original.

The archive should make the preserved information available to the user community, so they can determine the existence, description, location and availability of information stored in the OAIS. We include power users in the designated community, meaning that very large amounts of LIGO data can be analyzed: in section 3.5.3 we describe how this can be done at reasonable cost to the LIGO Laboratory.

It should be noted that LIGO has built a system that has successfully archived and preserved several petabytes of LIGO data. This system includes the following:

- Geographically distributed redundant data archive
- Grid enabled access
- High performance data transfer
- Periodic data archive technology migration
- Used by hundreds of scientists for LIGO analysis

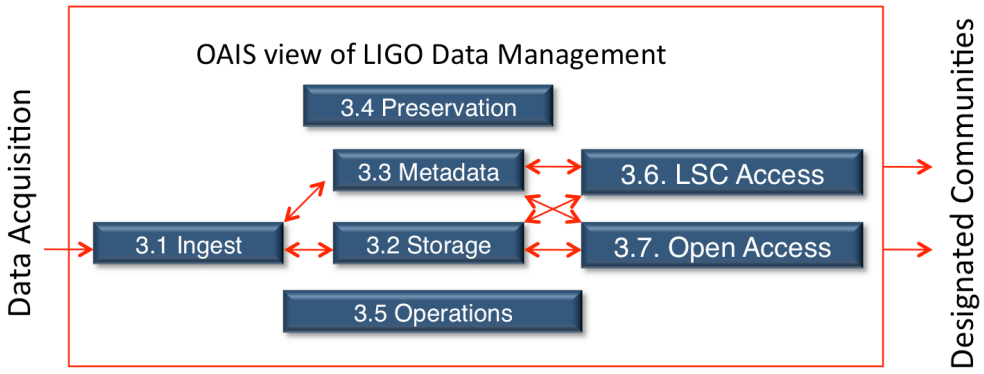


Figure 1: We factor the data management plan according to the OAIS (Open Archival Information System) model. Each functional component in this diagram is defined in the indicated section of this document.

1.5 Scope of Data Release

This plan covers data management in the LSC, LIGO Laboratory operations, and the Advanced LIGO Project, as well as incremental scope to support data release to the broader community; from the ingestion of the fresh data provided by the Advanced LIGO data acquisition systems (DAQ) to delivery to the designated communities of users.

"LIGO Data Access" means different things to different people. To be specific we consider three "designated communities" in the sense of OAIS as the users of LIGO data:

- LSC scientists, who are assumed to understand, or be responsible for, all the complex details of the LIGO data stream.
- External scientists, who are expected to understand general concepts, such as space-time coordinates, Fourier transforms and time-frequency plots, and have knowledge of programming and scientific data analysis. Many of these will be astronomers, but also include, for example, those interested in LIGO's environmental monitoring data.
- General public, the archive targeted to the general public, will require minimal science knowledge and little more computational expertise than how to use a web browser. We will also recommend or build tools to read LIGO data files into other applications.

The data delivery system will – as now – be built to give LSC scientists as much data access, computing, and bandwidth as is possible. In Phase 1 external scientists and the general public will have access to small data files and data tables, with web applications to show the meaning of these, and also some recommended software that can read the files. For External Scientists in Phase 2, we will provide the ability to access queries on metadata, triggers, etc., as well as providing significant quantities of data. LIGO computing will not be available to the broader community: they will need to provide their own resources for deep data mining or new analysis. However, an external scientist will be able to get large amounts of LIGO data as detailed in section 3.5.3 for analysis on national cyberinfrastructure facilities.

The LSC has already developed an outreach effort called the Einstein@Home project [8] to leverage an alternative distributed computing paradigm for its most formidable computing challenge, the search for gravitational waves from spinning neutron stars. This analysis puts reduced demand on quick turn-around and has low data flow, but requires petaflops of computing power. The analysis engine that underlies Einstein@Home utilizes much of the standard LSC software infrastructure described below; BOINC [9] is used to distribute work to thousands of volunteered personal computers and external computing grids world-wide.

1.6 Unified Data Management Plans

The LSC is supported through a combination of awards to individual investigators ("PI grants") and a central LIGO Laboratory Operations award. The NSF requires Data Management Plans for all awards [10]: *"All proposals must describe plans for data management and sharing of the products of research, or assert the absence of the need for such plans"*. We expect this Data Management Plan to be the basis for the plans of NSF-funded PI-grants, thus creating a collaboration in implementing components of this plan. We expect the results of computation on LIGO data to feedback, where appropriate, to create a higher quality LIGO archive for all, including both the LSC and the broader research community.

2 Advanced LIGO

2.1 Observing Runs

The four years during which Advanced LIGO was being installed and during the initial commissioning was an active time for the LIGO Data Management team, as they prepared to handle the new Advanced LIGO data stream, the anticipated detections, and Open Data OAIS paradigm (see Figure 1). To focus the attention of the team, there was a sequence of "engineering runs" every six months, in January and July. Observations with Advanced LIGO began in Sept 2015, and LIGO's first discovery followed soon; this first observing run finished in Jan 2016. The second observing run (O2) started Nov 2016, and will end in the third quarter of 2017. Figure 2 shows this timeline.

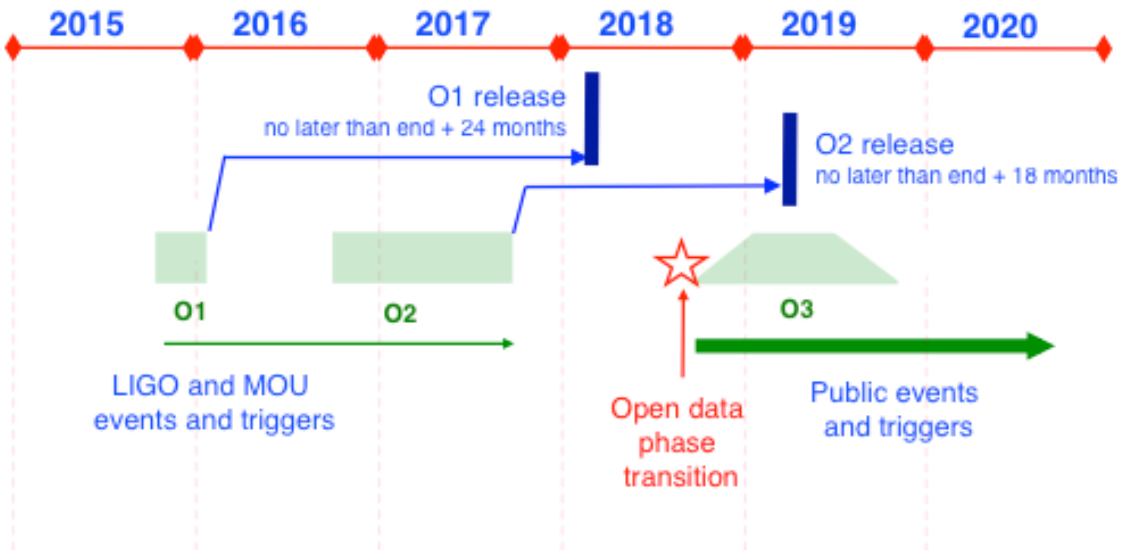


Figure 2: Schematic timeline for LIGO observations and transition to open data. Events and triggers are currently being issued to MOU observing partners for the electromagnetic follow-up program; these will be open to the public starting with O3. The LIGO Open Science Center (LOSC) is supporting the data releases. For more information see [23].

2.2 Timeline

The following table shows some of the milestones depicted in Figure 2:

Advanced LIGO Construction Milestones		
End of S6	2010 Q4	COMPLETE
Advanced LIGO Commissioning Starts	2013 Q1	COMPLETE
Three interferometers accepted	2015 Q2	COMPLETE
Community engagement for Open Data		
Broader Community Workshop 1	2011 Q4	COMPLETE
Broader Community Workshop 2	2014 Q4	DEFERRED
Observing Runs		
O1 start	2015 Q3	COMPLETE
O1 end	2016 Q1	COMPLETE
O2 start	2016 Q4	STARTED
O2 end	2017 Q3	FUTURE
O3 start	2019 Q1	APPROXIMATE

3 Work plan

We follow the OAIS functional model, with the following sections on Ingest, Storage, Metadata, Preservation, and Operations. Data Access is the sixth OAIS component, and it is covered in two sections: access for collaboration only (section 3.6) and broader community access (section 3.7).

As mentioned above, the plan outlined below is preliminary, and will be updated in future versions of this document. We expect implementation of the work plan will be divided among these resources:

- The LIGO Laboratory Maintenance and Operation Cooperative Agreement,
- The LIGO Data Grid operations grant for LSC access to data (refer to section 3.6)
- Possible additional funding to cover the new incremental scope for Open Data and Preservation – since this, as envisioned in section 3.7, is beyond the scope of current LIGO operations. An initial proposal to establish the LIGO Open Science Center (LOSC) was awarded by NSF in September 2012 for one year. Continued funding at a maintenance level for LOSC after 2013 has been under the LIGO Maintenance and Operations award.

For the software development work described in this plan, we will take an agile and incremental approach to delivery of tools. This means releasing software early and adaptively incorporating feedback from the user community. The lack of precedents to open data for gravitational wave astrophysics makes it difficult to predict what will be needed by outside researchers. Furthermore, full release of data with known signals will occur later in this decade and this plan will be refined as we gain greater experience with the broader research community.

3.1 Ingest

This data management plan covers the data products generated by the Advanced LIGO data acquisition (DAQ) systems, which are designated as the scientific data output of each instrument, and detector characterization metadata (data quality flags, calibration data, and the like). DAQ systems record their data in the International GW Interferometer data format called IGWD Frames. As was the case for Initial LIGO each interferometer (IFO) will record duplicate copies of the data, which will be written to independent file systems at each Observatory. Once these data are on stable storage they are declared acquired and shall be managed in perpetuity as described in this document.

The DAQ subsystem delivers “Validated Frame Data”. This means a well-formed, readable Frame file, with named, documented channels, properly identified by timestamp of its beginning and its duration. An additional part of the submission to the archive is the electronic control room logs.

For Advanced LIGO the data rates are about 25 megabyte/s per interferometer (IFO) for an aggregate rate of about 1.5 petabyte/year. The raw LIGO data are combined with auxiliary measurements and models to build a time series representing the gravitational-wave strain signal. This is then calibrated, and may also be flagged for quality control veto and/or cleaned. In addition, there are a large number of auxiliary instrumental and environmental monitoring channels that are also ingested.

LIGO data are stored in internationally standard data formats. This standard has been in place since 1997 as the consequence of a LIGO-Virgo agreement. It is the ‘Common Data Frame Format for Interferometric Gravitational Wave Detectors’ [11], also called an IGWD Frame file. This carries multi-channel time and frequency series, which can have different sampling rates for different channels. Broadly speaking, LIGO data currently consists of Frame files, each with thousands of channels, one of which represents the gravitational-wave strain, with all the other channels used for environment and instrument monitoring.

The information model of OAIS includes three packets: the Submission Information Packet (SIP) that is ingested, the Archive Information Packet (AIP) that is stored in the archive, and the Dissemination Information Packet (DIP) that is sent to the members of the designated community. Most LIGO data are expressed in Frames (supplemented by tables and real-time events see sections 3.6.1.2 and 3.7.2.2), and these will be used for the SIP, and AIP. While we expect most LSC scientists to continue choosing the IGWD Frame format, we will also disseminate the archive in other formats, such as HDF5.

3.2 Storage

Once the IFO data are acquired they are made available to the local data analysis facilities at each Observatory for low-latency analysis as well as distributed locations for coherent multi-detector analysis and to provide off-site storage redundancy.

A minimum of three full archival copies of the LIGO Data are being stored in at least three different geographic locations – Caltech, Hanford, Livingston – to provide access to the LSC and other users and to provide sufficient redundancy and security. LIGO will provide further data access to the LSC, as well as to the broader scientific community in the era of open data access. In addition, there will be remote data centers for geographically distributed access to public data.

We envision the data being stored (as now) in a distributed grid, and we will build delivery mechanisms that allow for large transactions, such as third-party transfers, and parallel data streams. Key services will be mirrored in at least two and preferably three geographically distributed sites. Initially we anticipate using the mirrors only for backup, but once the mirrors are in place we can use them for dynamic load balancing as well as automated failover.

Each local cluster must maintain a rapidly evolving set of middleware in as stable a fashion as possible, now called the LIGO Data Grid (LDG) server bundle; this software is rapidly evolving and requires effort to configure, support and maintain. The LDG currently uses the commercial Oracle HSM mass storage software from Oracle, commodity storage in the compute nodes, Linux based RAID servers, and the LIGO Data Replicator (LDR) to store and distribute data. LIGO IGWD Frame data are common to the majority of analysis pipelines, and so is distributed to all LDG centers in advance of job scheduling.

Data storage for LIGO is in a context of very large scale high-throughput computing. The Advanced LIGO instruments will generate over a petabyte of data per year. Even though only about 1% of these data are in the gravitational-wave strain channel (the rest consists of detector and environment monitoring information), LIGO data analysis is a formidable computing challenge. Binary inspiral, burst and stochastic searches can utilize many teraflops of computing power to analyze the data at the rate they are acquired, and optimal searches for periodic signals are computationally limited. LIGO's scientific pay-off is therefore bounded by the ability to perform computations on these data.

The LSC has converged on commodity clusters as the solution that meets its computational needs most cost effectively. LIGO has super-computer class requirements and data that can be handled efficiently in the simple parallel environment of clusters. In recent years the LSC has migrated to the grid concept of geographically distributed computing with clusters located at several sites. This approach has the advantage that it allows university researchers who are analyzing the data to contribute computational resources. Grid middleware allows for relatively easy access to data and computing power. If local resources are inadequate or a poor match a researcher can access additional resources on the grid.

3.3 Metadata

In addition to managing the data itself it is important to provide catalogs of what has been archived and where it is located. This applies to both the original IFO data and all derived data products. Therefore, we will provide a data discovery and cataloging tool based on the experience in Initial LIGO with the LIGO DiskcacheAPI tool, and the LIGO segment database.

The metadata includes:

- The database of when each detector was running and the global federation of segment databases for all GW observatories. This database will be used to carry data quality and injection information from multiple contributors. Monitoring and replication will ensure robustness of this database.
- Metadata and tools for the location and geometry of detectors, changing frame of reference, etc, etc. This would include ways to discover the existence of GW data through time and sky considerations.
- Data about injections in the data stream, so that these are not "discovered" by an open data user. Instructions must be forcibly clear about the existence of these injections in the LIGO data. Injections come in different kinds, corresponding to the search groups: compact binary coalescence (CBC), unmodelled Burst, continuous wave (pulsar), and stochastic (correlated).
- Data quality information indicating data that are not valid or which may not be suitable for some searches.

3.3.1 Calibration

Here we consider calibration information to be metadata. It is represented by a mathematical model of the instrument and calibration measurements which will also be stored, as part of the archive, as an additional IGWD Frame file. The calibration information may come later than the observational data, and calibrations may be revised and therefore versioned. We note that the strain channel is the only one that is guaranteed to be calibrated – not all of the numerous environmental monitoring channels. This task includes publishing response functions, noise estimates, autocorrelation estimates, and channel correlation so that significance of template matches can be assessed.

3.3.2 Data Services

Metadata services will be provided so that programs can read and write to the databases; then the visual clients (such as web forms) will be built on the services. In this way, users can use either a web form or a command-line or API mechanism to access the metadata.

3.4 Preservation

Independent of which data access era LIGO is operating in it is critical that all data generated by the LIGO instruments be properly managed to ensure their long-term viability and ease of access. In addition to the relatively small amount of Gravitational Wave (GW) data there is a wealth of auxiliary channels being recorded by the LIGO Observatories that are important to preserve and make available for analysis, e.g., studying historical trends from environmental monitors and removing instrumental artifacts from the GW data. Therefore, it will remain the policy of the LIGO Laboratory to retain, curate, and manage all recorded LIGO data in perpetuity.

3.4.1 Preserving the Bits

The data archive copies will be regularly scrubbed and repaired in the background for data corruption. This function will be integrated with periodic migrations to new storage technology.

Integrity checks will be done at several different layers within the storage and network hardware but there will also be application layer specific checks being provided. These will include independent hash verification for bulk transfers over the network, e.g., MD5, as well as internal checksums (CRC) stored within each Frame file. The internal file checksum provides the broadest coverage over the lifetime of the data as they are first calculated before the data are written to any storage device and are used to verify all data read back at analysis time by the standard IGWD Frame data I/O libraries.

3.4.2 Preserving the Meaning

There is a lot of documentation of LIGO resources that may be useful to the non-LSC scientist. This task is to collect and expose what is needed in a curated, coherent way, while ensuring that no information is made public that should not be. The task includes Getting Started, Tutorials, FAQ, References, Contributions, and Examples.

3.4.3 Data Curation and Publication

The moment of publication is the best time to expect the author of a journal article to spend time getting the content well-organized. The data-preservation community recognizes this, and suggests that relevant data be submitted along with the text of the article, and preserved with it. This can be the data files behind the tables and figures in the article, or it could be much more ambitious, including all that is needed to establish the results in the paper. The LSC is already adopting the former approach, using the LIGO Document Control Center as a release platform. This began in earnest in late 2011, as suggested in the National Science Board report [16].

3.5 Operations

3.5.1 LIGO Data Grid

The management of an organization where both its human and digital capital are widely distributed geographically presents unique challenges. The system of data and metadata production and dissemination will need resources to make sure it stays running, so that users are not seeing failed services, or mismatch between the data inventory and the actual delivery of that data. There are regular tests of services and what they produce, with downtimes posted in advance. Operations staff provide statistical analyses of the web log and service log data for input to quarterly and annual reports. These analyses include daily and monthly tracking of various LIGO/LSC sites and services, as well as providing aggregates and analyzing trends in the data.

Robust operation requires detailed problem tracking to ensure that services are maintained and that security issues are quickly and efficiently addressed; this includes a problem ticket system and monitoring services.

Building metadata services is a task that includes the status of clusters, services, jobs, and the connectivity between clusters, and is used to maximize the harvest of cluster cycle time.

3.5.2 Open Data Delivery

The scale of the LIGO Open Data Archive will be adjusted to meet the needs of the broad community of scientists: some small data with many users; and large data with few users. This task establishes the servers and disk cache to deliver LIGO data and its derived products in an efficient manner, a separation from private (LSC-only) data, maintaining digital object identifiers (“permanent URLs”) for data products by working with the library community.

3.5.3 Cloud Computing

This task allows users to compute on LIGO data, using their own facilities, national cyberinfrastructure (OSG, XSEDE, etc.), or other providers of wholesale computing (Amazon, Google, etc.). Users who want massive processing of LIGO data can obtain computer resources from the provider, and ask LIGO to help deliver the data there. The open data work package (section 3.7) will build the technical connections to enable this, and help such users to achieve their goals. This task enables data delivery; but LIGO itself is **not** supplying large scale computing resources for the broader research community beyond the LSC. The user may make a small number of large requests, or a large number of small requests; both situations have their own challenges for the data provider.

3.6 LSC Data Access and Computing

The objective here is to give the scientific collaboration (LSC) the data that they need for template matching and other advanced analysis: the whole dataset for massively parallel computing, or a small section of data; data filtered and calibrated in different ways and different versions. The data are not cleaned, but rather an annotation database keeps track of the different kinds of veto and data quality flags.

3.6.1 Data Access

We will support two kinds of data distribution: a bulk protocol based on files, and an application-based protocol based on data streams. For bulk data transfers, we envision a high performance file transfer protocol to distribute data stored in files. We will also build a stream-based way for applications to connect to a LIGO server, and have data channels delivered in near real-time as time or frequency series, with no significant file storage by the client. We expect to build on current technology: the LIGO Data Replicator (LDR) for file transfers, and the Network Data Server (NDS), for streaming data.

3.6.1.1 Streamlined Data Files

The open data infrastructure will be able to deliver streamlined data files that are easier to use for most open data users, for example HDF5 in addition to the existing LIGO Frame format. Instead of the current paradigm of segments and fixed frames, the system will be able to deliver LIGO strain data in a single file over an arbitrary interval, annotated with data quality and injection flags.

3.6.1.2 Events and triggers

The rise of new synoptic (repeating) sky surveys has seen the establishment of the VOEvent format [12] for reporting and rapid follow-up of astronomical transients. We propose to facilitate this science by operating LIGO services that allow publication of gravitational-wave events and receipt by subscribers within seconds. Subscribers may selectively subscribe to event feeds, query past events, and investigate possible event correlations (e.g., GRBs and neutrino events).

We will continue and refine our program of real-time alerts for relevant astrophysical events, both to and from LIGO. From the emerging VOEvent network, we will utilize external triggers for searching LIGO data, and we will build a stream of high-significance LIGO triggers for immediate dissemination to external observatories. Some 64 groups are currently signed up to follow LVC alerts, with an open invitation[1] to the world's astronomers to participate in the rapid search for the counterpart. This maximizes the chance of immediate identification of multi-messenger events, and the consequent scientific payoff from LIGO. The system will be based on the highly successful GCN program that NASA has operated for 20 years. As with other LIGO data, event release will be restricted in Phase 1 to LSC and other entities with which there is an MOU, and then opened more broadly in Phase 2 as described in section 1.2 above.

In addition to real-time event exchange, there will be an archive of events; it will be distributed to insure data integrity and easy failover, with designated individuals to move the data into the archive. As part of this project, database solutions that allow event storage and annotation will be explored.

3.6.2 Software

3.6.2.1 Open Source Analysis Libraries

LALSuite consists of extensive libraries in Python and C for LIGO data analysis. Matapps will connect to the Matlab signal processing toolkit for LIGO search pipelines. This task will also support visualization software. The code will be made into a collection of independent modules, including, for example: power

[1] http://www.ligo.org/science/GWEMalerts.php

spectral density, masking bad data, band limiting, trends and non-gaussianity, barycentering and Doppler correcting, space/time frames.

3.6.2.2 Computing

For computing across the LIGO Data Grid (LDG), we envision tools such as Globus for initiating jobs, Pegasus for workflows, and Condor for job queue management. As the analysis workflows for this new branch of astronomy are evolving rapidly, significant effort is required to work closely with the upstream middleware development team to ensure efficient use of the LDG clusters. However, these are complex software packages, and staff will be required to offer performance tuning, job scheduling efficiencies, memory utilization, file management, and general debugging support for intermittent job failures. There will also be a need to maintain and administer the virtual organization (accounts, access etc.).

3.6.3 Help and Documentation

3.6.3.1 User Services

For users to extract maximum science from the LIGO data stream, they will need help at many levels, including a Forum for discussions with LSC and other scientists, and a way to submit help tickets that will be answered rapidly. Responses will be fed back into a searchable knowledge base, so that the next time a user has the same question, they will not need the helpdesk, but can find the answer in the knowledge base.

We will also log and measure the uptake and usage of the data. This implies the gathering and summarization of such information from server logs, to give an accurate assessment of user behavior.

3.6.4 Management

We expect much of the development and hosting of services for the open data initiative to take place within the LIGO Lab, but with significant contributions from LSC institutions, through the LIGO Open Science Center (LOSC). Those contributions may be formal subcontracts for LOSC software development, or may be informal ‘contributed software’. The LIGO Lab will provide leadership and coordination for both in-Lab and LSC contributions. Progress will be reported through the annual NSF review of the LIGO Lab.

We will leverage and extend the existing identity management system that LIGO is developing, to ensure that data, services, and software remain properly contained. The beginning of Phase 2 will mean that new data is still private to the LSC, but will become open after the proprietary period. The task here is to keep up to date with security solutions, and also verify and approve authorized requests.

3.7 Public Open Data Access

In September 2012, the LIGO Open Science Center (LOSC) was started with a 1-year award from NSF. LOSC is a fabric to enable robust scientific discourse about the LIGO data products for somebody *who has never met any LSC member*. This document elaborates the following six work packages that we feel cover the NSF requirement to make LIGO data available and useable by the broader community: Data Services, Data with Quality Information, Interoperability, Community, and Software.

We take the “broader community” to include future scientists, hence the Data Preservation principle. Each of these principles drives a work package, as listed below. A seventh task is engineering, so that everything actually happens efficiently and reliably. The principles are explained below, with the required work packages.

Public Open Data Phase 2 will supply several categories of product: Real-time event distribution, calibrated strain data, data quality information, catalogs of data products, and other metadata. Users will be able to get data products through web services ("code") as well as web interfaces ("click"). There will also be a tutorial and software collection, as elaborated in the next section.

3.7.1 Data Services

LOSC will work with the LIGO Data Grid (LDG) for efficient supply of data and real-time events, delivering h[t], annotation, and other channels for the LSC and the broader community. Additional products may include barycentered h(t), h(omega) for given sky position, h[t] products optimized for search type, and selected PEM channels. These can be achieved through a workflow of filters on the data, combined with a query on the annotation to maximize the amount of useful and clean data for a given application.

For open data release, other formats may be more useful than the IGWD Frame file. The LOSC open data releases have also provided the HDF5 format, an internationally recognized standard that will continue to be readable far into the future. Thus we do not provide software to handle IGWD Frames, that must run on "all" machines that users may have, rather we leverage the large amount of software already developed for HDF5.

3.7.1.1 LOSC Web Services

We are working with the LDG to build services and other access to h[t], the annotation database, its queries and responses, and documentation to make LIGO data accessible to the broader community. An important principle here is to provide as much as possible as both "click" and "code", so that a human can use a browser ("click") to discover and use the data, and in addition, a program can be written ("code") to fetch and use the data.

3.7.1.2 Real-time Events

A primary purpose of the Open Data program is to expand the rapid follow-up program for possible GW signals. This task is to build an infrastructure for rapid delivery of events, with associated data, including classification, and selection on that data. If robotic telescopes are to slew to the trigger, users will want careful selection. This task also includes computing and understanding the LIGO signal associated with external triggers (e.g. gamma-ray burst observatories).

We have deployed a system that uses the Virtual Observatory VOEvent Transport Protocol (VTP) [17], to communicate real-time alerts to follow-up observers. Many of these have studied gamma-ray bursts since 1997, and we intend that they can easily participate in the search for GW afterglows, without need to retool the software. LIGO has established a working relationship with the NASA GCN system, that has distributed these rapid alerts for NASA since the 1990's, to find gamma-ray bursts and other high-energy phenomena. LIGO has put in place a successful plan to engage the astronomical community (see section 3.7.1).

3.7.2 Data with Quality Information

The LIGO data stream, like that from any instrument, has artifacts that come in a wide variety of forms, and we expect to deliver the nature and location of these artifacts along with the gravitational wave signal. The team involved in this task must characterize the types of glitches coming from aLIGO, and build ways to eliminate them or flag them in the open data product. The task includes collecting and maintaining data quality and injection metadata (Boolean flags sampled at 1 Hz), to enable the definition and production of a variety of LIGO data products. The LOSC work will streamline and simplify that metadata for use by the broader community.

3.7.2.1 Building Data Quality Information

Annotating data with data quality (DQ) information allows interoperation of the results of many pipelines and detector characterization and other mining software. Part of the mission of LOSC will be working with those who know the data – and their software pipelines – to add those "opinions" on the nature of LIGO data, so the informal knowledge in their heads is collected and preserved for future users.

3.7.2.2 Event Repository

In addition to rapid follow-up, users will want to study collections of past triggers, looking for correlation with astronomical and artifact streams, and correlation with other (non-LIGO) event streams.

3.7.3 Interoperability

Open LIGO data products and services should work well with scientifically relevant non-LIGO systems. This task will ensure that LIGO open data works well with normal tools of astronomy and of signal processing. This task will position LIGO as a vigorous player in relevant international standards.

3.7.3.1 Standard Formats

We will present data to the broader community with standard formats, for which standard tools are available, such as HDF5 and VOEvent, and other virtual observatory standards. This follows the recommendations in the National Science Board report, December 2011 [16].

3.7.3.2 Real-time Events

We expect to use the international standard VOEvent for real-time messages to the astronomical community, to get rapid follow-up of LIGO triggers from the widest variety of telescopes.

3.7.4 Community

We do not wish to develop closed, inflexible requirements and then build to them whether wanted or not. Rather, we will start with a process of community engagement, discovering what is wanted from the LIGO data stream, by professional astronomers, instrument specialists of other observatories, amateur astronomers, and the interested public. We began with a survey of astronomers and a workshop to generate requirements, in October 2011: for data products and data quality, for software and web-services, and for user-facing services. This task involves understanding community needs, enabling the community to work together, and responding to their questions.

In terms of OAIS, this is about identifying a "Designated Community" (DC) for the receipt of LIGO data; however we will have several DC's: (1) LSC scientist, (2) non-LSC scientist, and (3) public (EPO) access. The community of (1) is already well-engaged – see section 1.1 – and the task of the open data initiative is to extend access to communities (2) and (3).

We have begun a conversation that will continue throughout the LIGO venture, talking to real users of LIGO data from all corners of the GW community, understanding the nature of the Designated User Communities and their expectations and requirements. In order to evolve into the most useful facility it can be, the LIGO data facility must proactively pursue community feedback from all the designated communities to collect ideas for improvements, broader suggestions of new features and totally new tools and services. This active pursuit of feedback from the LIGO user community will be woven into our community engagement activities so that advocacy becomes a two-way flow of information, showing the community what we have while also listening to the community.

In building the LIGO Tool Box, we want to involve the LSC community in contributing modules, and yet those modules should conform to best practices. We expect software to have levels of review: at the lowest level that which is contributed but not reviewed; at the next level that which is reviewed, tested, and packaged; and at the highest level that which has been through stringent review and used for a peer-reviewed LSC publication.

3.7.4.1 Help Desk and Forum

For users to extract maximum science from the LIGO data stream, they will need help at many levels, including a Forum for discussions with LSC and other scientists, and a way to submit help tickets that will be answered rapidly.

3.7.4.2 Open Data Workshops

We are consulting with the broader community on how to maximize the usefulness of the LIGO Open Data Archive. We began the process in summer 2011, and continue consultation, workshops and events right through full open-data release and beyond.

3.7.4.3 Logging and Assessment

The real story of what users want is contained in the services and data that they utilize. This task is the gathering of such information from server logs, to give an accurate assessment of user behavior.

3.7.4.4 Outreach with Open Data

Announcement of the first discovery was an exciting time, and interest in all "black-hole" things was at an all-time high. However, the physics and the instrument are difficult to grasp for most people, and a fresh effort, based on Open Data, perhaps in conjunction with a prominent Arts institution, could make LIGO be exciting for a lot of people. This will build on existing outreach efforts within the LIGO Laboratory and the LSC.

3.7.5 Software

LIGO has been building analysis pipelines and associated libraries for a decade, and these are mostly open source, so there is no need to have these "made open". LOSC will provide significant documentation and framing to make it easy for the broader community to take advantage of these. However, LOSC will also provide a certain amount of software built specifically for the broader community.

3.7.5.1 Software for Open Data

We will provide I/O libraries for whatever formats LOSC uses to deliver data and metadata. This project will emphasize interoperability, providing command line, API, and web form access, services as well as installed software. Another emphasis will be on gluing and framing with existing software products rather than building from scratch ("*use plugin X to import the image into Photoshop")*. Community input will be paramount here in creating the precise work package. The task will also include a new package for showing time-frequency and correlations that has ultra-simple installation (app or applet), and is open source. Specification and software will depend on community feedback.

The released software products may include: data file reader; tools for Short Fourier Transforms (SFT); display and GUI in time and/or frequency, such as that used now for the control-room display; GUI and API for databases and metadata; computing power spectral density; tools to mask bad data; band limiting and heterodyning; searching for non-Gaussianity; computing trends in the data; tools for barycentering; and manipulating celestial coordinates and time frames. We will also build a "software forge" so that contributed software from the LSC can be made available to the broader community, after suitable packaging and review.

4 References

[1] A. Lazzarini, S. Whitcomb, J. Marx, R. Weiss, D. Reitze, B. Barish, P. Saulson, A proposal for providing open access of LIGO data to the broader research community, https://dcc.ligo.org/LIGO-M080072/public/main

[2] Open Archival Information System http://en.wikipedia.org/wiki/Open_Archival_Information_System

[3] The Open Archival Information System: Introductory Guide, Brian Lavoie, at http://public.ccsds.org/publications/archive/650x0b1.PDF

[4] How to join the LSC: http://www.ligo.org/about/join.php

[5] GCN: The Gamma-ray bursts Coordinates Network, NASA http://gcn.gsfc.nasa.gov/

[6] LIGO Scientific Collaboration, Predictions for the Rates of Compact Binary Coalescences Observable by Ground-based Gravitational-wave Detectors, Class. Quant. Grav. 27: 173001 (2010)
[7] An Astrophysical Metric for LIGO Open Data Release, S. Fairhurst, I. Mandel, A.Weinstein, V. Kalogera
https://dcc.ligo.org/LIGO-T1000414/public/main
[8] Einstein@home: http://einstein.phys.uwm.edu/
[9] BOINC, open-source software for volunteer computing, http://boinc.berkeley.edu
[10] National Science Foundation Data Management Plan Requirements
http://www.nsf.gov/bfa/dias/policy/dmp.jsp
[11] Common Data Frame Format for Interferometric Gravitational Wave Detectors
https://dcc.ligo.org/LIGO-T970130/public/main
[12] Sky Event Reporting Metadata (VOEvent), Rob Seaman, Roy Williams,
http://www.ivoa.net/Documents/latest/VOEvent.html
[13] US Virtual Astronomical Observatory, http://usvao.org
[14] LIGO Data Analysis System, http://www.ldas-sw.ligo.caltech.edu/cgi-bin/index.cgi
[15] Ligotools, http://www.ldas-sw.ligo.caltech.edu/ligotools/
[16] National Science Board paper NSB-11-79 : Digital Research Data Sharing and Management, December 2011,
http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf
[17] GCN: The Gamma-ray Coordinates Network, http://gcn.gsfc.nasa.gov/
[18] Lambda, NASA's data center for Cosmic Microwave Background (CMB) research
http://lambda.gsfc.nasa.gov/
[19] NASA/IPAC Infrared Science Archive, http://irsa.ipac.caltech.edu/
[20] High Energy Astrophysics Science Archive Research Center (HEASARC)
http://heasarc.gsfc.nasa.gov/
[21] M. Vallisneri, J. Kanner, R. Williams, A. Weinstein, B. Stephens, proceedings of the 10th LISA Symposium; see http://losc.ligo.org for the S5 data release and more information about the LIGO Open Science Center, also http://arxiv.org/abs/1410.4839
[22] Binary Black Hole Mergers in the first Advanced LIGO Observing Run,
http://arxiv.org/pdf/1606.04856.pdf
[23] Prospects for Observing and Localizing Gravitational-Wave Transients with Advanced LIGO and Advanced Virgo,
https://arxiv.org/abs/1304.0670
[24] Updated O2 observing run schedule,
https://dcc.ligo.org/LIGO-L1700028/public

5 List of Acronyms

AIP	Archive Information Packet
aLIGO	Advanced LIGO
API	Application Programming Interface
DAQ	Data Acquisition
DC	Designated Community
DIP	Dissemination Information Packet
DMP	Data Management Plan
EPO	Education and Public Outreach
GCN	Gamma-ray Coordinates Network
GRB	Gamma Ray Burst

GW Gravitational Wave
HDF5 Hierarchical Data File format, version 5
IAU International Astronomical Union
IFO Interferometer
IGWD International Gravitational Wave Detector
LAL LIGO Algorithm Library
LDAS LIGO Data Analysis System
LDR LIGO Data Replicator
LIGO Laser Interferometric Gravitational wave Observatory
LOOC-UP Locating and Observing Optical Counterparts to Unmodeled Pulses
LSC LIGO Scientific Collaboration
LSST Large Synoptic Survey Telescope
LVC LIGO-Virgo Collaborations
MOU Memorandum of Understanding
NDS Network Data Server
NSF National Science Foundation
OAIS Open Archival Information System
PEM Physical and Environmental Monitoring
PI Principal Investigator
RLS Replica Location Service
SIP Submission Information Packet
VAO Virtual Astronomical Observatory
VO Virtual Observatory

Appendix A: update January 2012

This LIGO Data Management Plan (DMP) is intended to be a 'living' document, updated every year to reflect the dynamic landscape of data management and open data. This Appendix is the new part of the -DMP.

5.1 A.1 Progress during 2011

A.1.1 User requirements much better defined

As a result of a survey of the broader community (80 responses), and a workshop (40 attendees), it is becoming clear that astronomers are very excited about the upcoming release of LIGO data. For a full report on what has been learned, see "Community Input for LIGO Open Data": https://dcc.ligo.org/LIGO-P1100182/public/

A.1.2 LOSC proposal submitted

A proposal has been submitted (Nov 2011) for the LIGO Open Science Center (LOSC), giving access to LIGO data to a broad user community. The LOSC will gather relevant data products generated by the LIGO Laboratory and the LIGO Scientific Collaboration (LSC), and make those data products available to outside users along with tools and support for those users' goals. The LOSC will operate before, during, and beyond operation of the Advanced LIGO observations, providing user services that are not funded or available in any other way. The LOSC will provide: (a) a web portal, client tools, and a layer of data access services through which the user can discover and access a range of LIGO data and metadata products; (b) long-term data curation; (c) a knowledge-base "LIGOpedia", helpdesk, online tutorials, and user feedback forums; (d) sponsorship of tutorials, workshops and conferences to help users derive maximal science from the data and communicate their experiences and findings to each other and to the

LSC; and (e) LIGO data-centric resources to facilitate LIGO and LSC education and public outreach efforts about gravitational wave and related science.

A.1.3 Data publication with article

LIGO and the LSC have begun a program of data publication with each journal article. The data files for each figure and table are uploaded along with the text, using the LIGO Document Control Center. See section 3.4.3 for more information.

5.2 A.2 Plans for 2012

A.2.1 Engineering runs now defined

In advance of Advanced LIGO, Engineering Runs will happen each January and July, starting in 2012. The first will concentrate on low-latency delivery of data to the computing pipelines. It is expected that the synthesized data from these runs will be available to the broader community in later engineering runs.

A.2.2 Trigger Delivery

Testing and prototyping of delivery of LIGO triggers to the public is expected to be through some combination of VOEvent, Skyalert, and NASA's GCN system.

A.2.3 Second Open Data Workshop

LIGO plans a second Gravitational Wave Open Data Workshop in the second half of 2012, which will include tutorial aspects and hands-on training.

5.3 A.3 Changes to Data Management Plan

Changes from the version of Jan 2011 are:

- Software Engineering Runs are now called Engineering Runs. Also the schedule of these has crystallized, and those new projected dates are in the body of the DMP, and depicted in the schematic timeline (Figure 2).
- A bullet has been deleted about using PI-grants to implement components of the plan (beginning of section 3). Rather it is expected that the work will be carried out by the remaining three efforts listed there.
- A section has been removed entitled "Computing to Support LIGO Science", with its words distributed elsewhere.
- Section 3.3.2 'Annotation Database' has been removed. This will be carried out by the LIGO Lab and LSC, and does not need to be called out in the DMP.
- Section 3.4.3 has been added, regarding publication of data with each journal article.
- Section 3.6.1.2 has been modified, to explicitly call out the NASA GCN (Gamma-ray Coordinates Network) as a mechanism for distributing LIGO triggers.
- Section 3.6.4. Channel Registry, has been deleted. This will be carried out by the LIGO Lab and LSC, and does not need to be called out in the DMP.
- The word 'openLIGO' has been replaced everywhere with LOSC, the LIGO Open Science Center.
- Section 3.7.2 on 'useful' data now refers to 'data quality', a name in line with the rest of the GW community.

Appendix B: update January 2013

This LIGO Data Management Plan (DMP) is a 'living' document, updated every year to reflect the dynamic landscape of data management and open data. This Appendix serves as progress report on implementation, and also noting actual changes to the text of the DMP.

5.4 B.1 Progress during 2012

B.1.1 LOSC awarded and started

The NSF award for the LIGO Open Science Center (LOSC), has allowed the formation of the open data team and the start of the implementation. Data archive products are being computed for the S5 and S6 runs of initial LIGO, although these are currently restricted to LSC member access only.

B.1.2 Data release for GRB051103

The LSC has released a comprehensive dataset covering the 'significant non-detection' of a gamma-ray burst in November 2003. The burst was thought to come from a nearby galaxy, in which case LIGO might have detected gravitational waves from the same source; we infer from the absence of such a detection that that the source is much further (i.e. behind) that nearby galaxy. The data release is at http://www.ligo.org/science/GRB051103/index.php

B.1.3 Website prototype

The LOSC team has built a website for open data, although it is still private and "LSC only". The site is at https://losc.ligo.org, and offers some tools for evaluating which detectors were operating at which times in the past, as well as some HDF5 files containing S5 and S6 data, complete with data quality information.

5.5 B.2 Plans for 2013

B.2.1 Building the LOSC archive

We will continue the prototyping and testing of a web-based archive for the (past) S5 and S6 runs, as well as being involved in the planning of the Advanced LIGO data management, so that future science runs will fit into the same archive structure.

B.2.2 Engineering runs underway

Prior to the first Advanced LIGO science run, ongoing Engineering Runs will happen each January and July, starting in 2012. The third run (ER3) is now underway, prototyping the data management for aLIGO, especially low-latency data transfer.

B.2.3 Second Open Data Workshop

LIGO plans a second Gravitational Wave Open Data Workshop in the second half of 2013, which will include tutorial aspects and hands-on training.

5.6 B.3 Changes to Data Management Plan

Changes from the version of Jan 2012 are:

- Changing the words from those indicating future milestones to those indicating that these have happened.
- Better definition of the archive data products that LOSC will offer to the broader community.

Appendix C: update February 2014

This LIGO Data Management Plan (DMP) is a 'living' document, updated every year to reflect the dynamic landscape of data management and open data. This Appendix serves as progress report on implementation, and also noting actual changes to the text of the DMP.

5.7 C.1 Progress during 2013

C.1.1 S5 open data is near release

During 2013, the LIGO Open Science Center has flourished, with funding from NSF, producing keystone technologies and prototypes that will service a large number of users once GW detections have started. In preparation for this two things have happened: The LSC Council has resolved to release the initial LIGO science run (S5, 2005-2007); and the LOSC has prepared a web-based data warehouse for the gravitational wave channel at 4 kHz, the data-quality tests at 1 Hz, and the injections at 1 Hz, and additional derived quantities about the data. Web based visualization shows the quality metadata at different timescales. Also in the release are the hardware injections that were both noted at the time, and recovered by the LOSC team; this is to prevent "discoveries" by open-data users, which are actually known injections. LOSC has built a web presence at http://losc.ligo.org, which will have significant content, once LSC Review is complete and S5 is released. The LSC-private version of the LOSC S5 release is at http://losc-dev.ligo.org.

C.1.2 Engineering Runs continue

All of the rapid-event data release protocols are being exercised during the engineering runs, which have been following a 6-month schedule as indicted in Figure 1.

C.1.3 Rapid alerts and observational follow-up of LIGO alerts

We have deployed a system to communicate real-time alerts to follow-up observers. Many of these have studied gamma-ray bursts since 1997, and we intend that they can easily participate in the search for GW afterglows in the late 2010s. LIGO has established a working relationship with the NASA GCN system, that has distributed these rapid alerts for NASA since the 1990's -- to find gamma-ray bursts and other high-energy phenomena. LIGO has set up a plan to engage the astronomical community, with an open invitation to the world's astronomers to participate in the rapid search for the counterpart. Some 64 groups are signed up (February 2014) to follow LVC alerts.

5.8 C.2 Plans for 2014

C.2.1 Review and release data from initial LIGO (S5 run, 2005-2007)

C.2.2 Prepare S6 for release pending feedback from the S5 release

C.2.3 Webinar tutorials about LOSC data products

C.2.4 Prototyping the Rapid Follow-up system in Engineering Runs

Appendix D: update January 2015

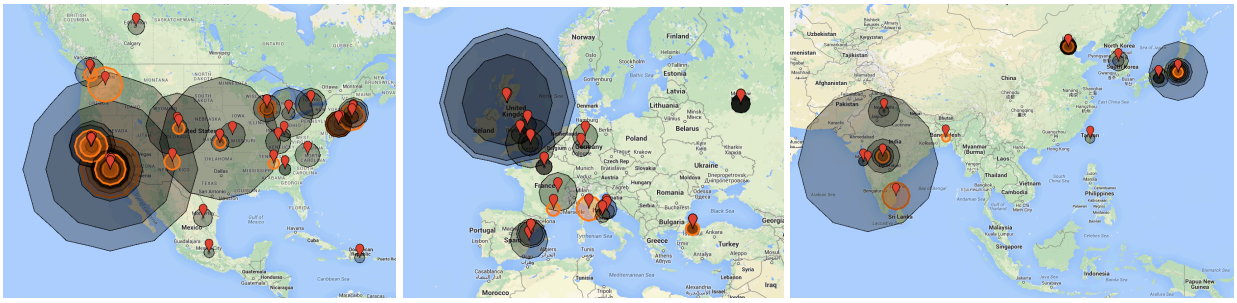
This LIGO Data Management Plan (DMP) is a 'living' document, updated every year to reflect the dynamic landscape of data management and open data. This Appendix serves as progress report on implementation, and also noting actual changes to the text of the DMP.

Release of the Initial LIGO datasets (S5 already and S6 being considered) provide a prototype and infrastructure for future release of LIGO open data, putting the project in a good position for future obligations of open release of Advanced LIGO data.

5.9 D.1 Progress during 2014

D.1.1 S5 open data is released

During 2014, the LIGO Open Science Center has built a web presence at http://losc.ligo.org, which now has terabytes of significant content from the S5 run, all reviewed and approved by the LSC. More information is available at "The LIGO Open Science Center", by M. Vallisneri, J. Kanner, R. Williams, A. Weinstein, and B. Stephens (http://arxiv.org/abs/1410.4839).



The LOSC website collects analytics information, and these three images summarize the global reach. Each pin represents an IP-address that has downloaded S5 data files, each being ~120 Mbyte covering a 4096-second interval, and the area of the disk around the pin is proportional to the number of files downloaded between the S5 release in September and the end of 2014, the total being 17801, excluding Googlebots.

D.1.2 Future LIGO Runs

Engineering runs continue, with the completion in December 2014 of the ER6 run, and the expectation of ER7 in Q2 of 2015. A science run of Advanced LIGO called O1 is expected in 2015Q3.

5.10 D.2 Plans for 2015

D.2.1 Review and release data from S6

In the first half of 2015, we expect to release the S6 data from Enhanced LIGO (taken 2009-2010), once the LSC review is complete and the LSC agreement on hand. The LSC-private version of the LOSC S6 release is already available at http://losc-dev.ligo.org.

D.2.2 Prototyping release of Advanced LIGO data

The LIGO Open Science Center is following closely the Advanced LIGO data formats and distribution, prototyping the open release to be ready when the time comes.

Appendix E: update August 2016

This LIGO Data Management Plan (DMP) is a 'living' document, updated every year to reflect the dynamic landscape of data management and open data. This Appendix serves as progress report on implementation, and also noting actual changes to the text of the DMP.

5.11 E.1 Progress during 2015

- The "S6" data from Enhanced LIGO (2009-2010), has been released, similar to the S5 release noted above, with the details available at https://losc.ligo.org/S6
- Advanced LIGO was in observational mode (the "O1" run) during the last quarter of the year. The data are still being analyzed at the time of this writing.

- Gravitational waves have been detected (Sept 2015), making the open data trigger of "plentiful detections" seem much closer.
- LOSC has built three data releases:
    - for the GW150914 event, available at http://dx.doi.org/10.7935/K5MW2F23,
    - for the GW151226 event, available at http://dx.doi.org/10.7935/K5H41PBP,
    - for the LVT151012 event, available at http://dx.doi.org/10.7935/K5CC0XMZ

5.12 E.2 Plans for 2016

- LOSC will build a "Data Release" for any detected and published events, as defined in section 1.3.2. There will be about an hour of data around each event.
- The data release for published events will include not just strain data, but machine-readable versions of the published tables and figures.

Appendix F: update June 2017

This LIGO Data Management Plan (DMP) is a 'living' document, updated every year to reflect the dynamic landscape of data management and open data. This Appendix serves as progress report on implementation, and also noting actual changes to the text of the DMP.

5.13 F.1 Progress since Aug 2016

- The "O1" data release from Advanced LIGO, has been built and is being reviewed.
- Advanced LIGO was in observational mode (the "O2" run) from Nov 30, 2016. The data are still being analyzed at the time of this writing.
- A "catalog" infrastructure has been prototyped, to deal with plentiful events, providing catalog, web page, and various plots.
- LOSC has released another binary black hole event:
    - GW170104 available at https://doi.org/10.7935/K53X84K2

5.14 F.2 Future Plans

- LOSC will build a "Data Release" for any detected and published events, as defined in Section 1.3.2. There will be about an hour of data around each event.
- The data release for published events will include not just strain data, but also machine-readable versions of published tables and figures.
- LOSC will release the full data from the O1 run no later than Jan 2018, 2 years after the end of that run.
- LOSC will release the full data from the O2 run no later than 18 months after the end of that run
- Beginning with O3, the era of Open Data will start, as defined in this document, including rapid public alerts of significant triggers being found; and begin the era of regular 6-month public data releases after an 18-month proprietary period.