An Open Data Policy for LIGO

A. Lazzarini

Caltech LIGO Laboratory

LSC-Virgo Meeting
Amsterdam, The Netherlands
22 - 25 September 2008

LIGO-G080445-00-M

LIGO

At the June 2008 meeting in Orsay I reported:

LIGO

Other news:

- Also part of the approval resolution …

. . .

WHEREAS, the NSF management shall report back to the National Science Board within 12 months of this award what efforts have been undertaken and what provisions have been implemented to make the data obtained under this award available and useable to the broader research community;

. . .

- This is a directive to the NSF (not LIGO or the LSC at this point)
- The model under which we launched the LSC in 1996 may no longer be a valid one in 2009+ …
- LIGO Directorate are considering options prior to NSF making a request to the LIGO Laboratory
  - Take the initiative to consider what the Lab and the LSC would like to see happen given that there are changes coming …
  - There will be plenty of time to discuss within the LSC as strawman options come into sharper focus …

In the ensuing period of time:

- The Directorate met with (i) the former Spokesmen, (ii) former Laboratory Director and Deputy Director to develop a strawman proposal for consideration by the broader LSC
- This committee:
  - Consulted with astrophysicists who were knowledgeable about LIGO and the LSC;
  - Consulted with several [NASA] astrophysical observatories and data centers with open data policies;
  - Consulted with LSC working group chairs with data analysis responsibilities
  - Numerous interactions as a committee to finalize the draft proposal for circulation to the LSC and to Virgo leadership

Context: The Big(ger) Picture

- Public release of data from government financed projects is a mandate coming from the U.S. Congress
  - General recognition that many of these databases constitute a rich treasure that must be protected, preserved, exploited
  - Substantial public funds have been invested in generating the data
- Experience from other fields that have profited from common data formats
  - Earth observations, climate modeling, seismology;
  - Syntheses of cross-disciplinary databases have enabled new analyses, discoveries through data mining.
- NSF policy already calls for data sharing …

LIGO

Source: NSF Grant Policy Manual

734 Dissemination and Sharing of Research Results

a. Investigators are expected to promptly prepare and submit for publication, with authorship that accurately reflects the contributions of those involved, all significant findings from work conducted under NSF grants. Grantees are expected to permit and encourage such publication by those actually performing that work, unless a grantee intends to publish or disseminate such findings itself.

b. Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. Privileged or confidential information should be released only in a form that protects the privacy of individuals and subjects involved. General adjustments and, where essential, exceptions to this sharing expectation may be specified by the funding NSF Program or Division for a particular field or discipline to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate the legitimate interest of investigators. A grantee or investigator also may request a particular adjustment or exception from the cognizant NSF Program Officer.

c. Investigators and grantees are encouraged to share software and inventions created under the grant or otherwise make them or their products widely available and usable.

d. NSF normally allows grantees to retain principal legal rights to intellectual property developed under NSF grants to provide incentives for development and dissemination of inventions, software and publications that can enhance their usefulness, accessibility and upkeep. Such incentives do not, however, reduce the responsibility that investigators and organizations have as members of the scientific and engineering community, to make results, data and collections available to other researchers.

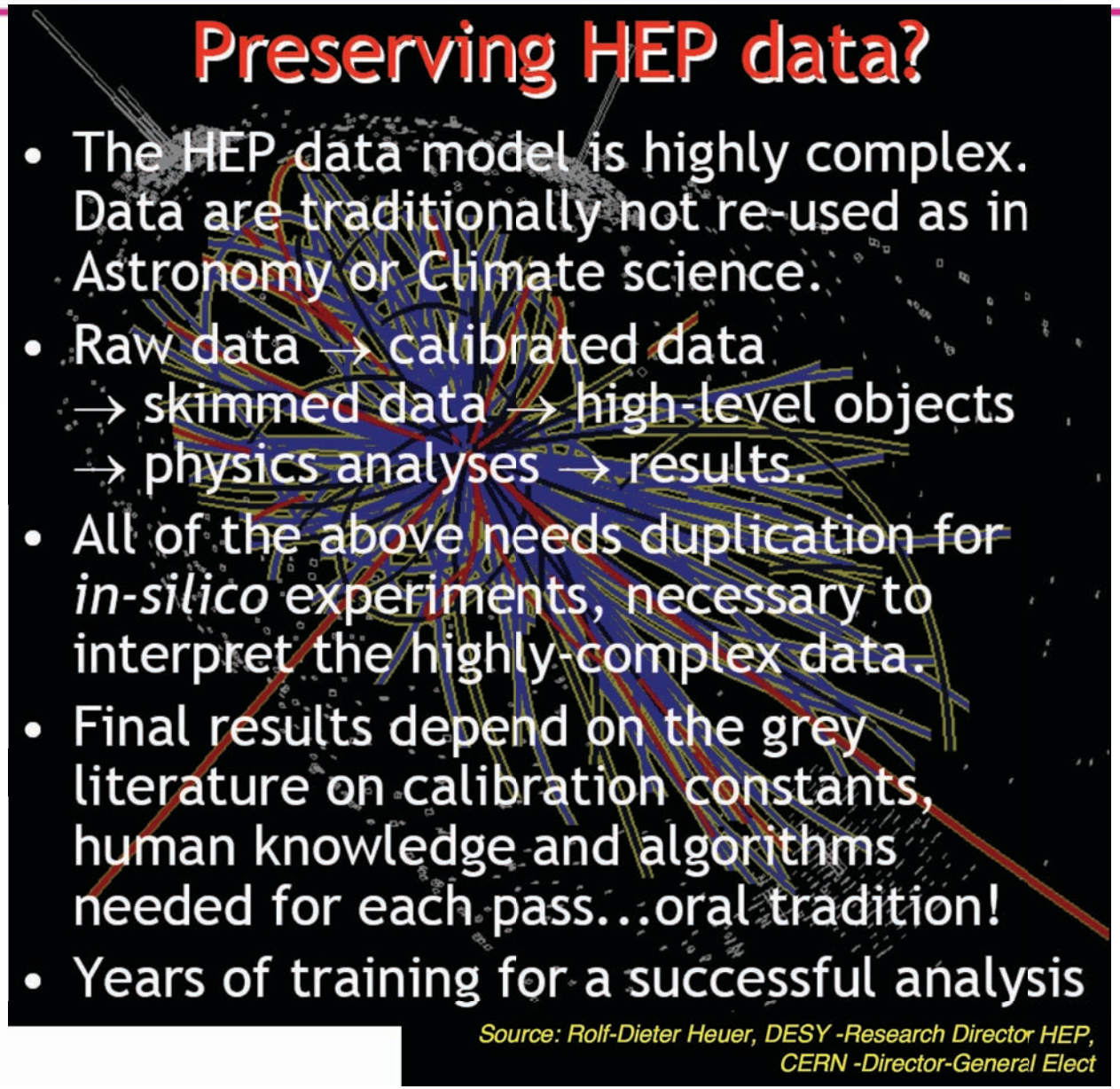
Why LIGO? Why Now?

- Issue of open access to results of publicly funded research is making its way to the forefront of deliberations on funding new & continuing programs

- EU (CERN/LHC):
  - Open Access --
    - Ref: Conference on Academic Publishing in Europe
      Rolf-Dieter Heuer, DESY -Research Director HEP, CERN -Director-General Elect -, Berlin, Jan.2008.
      - http://library.desy.de/sites/site_library/content/e14/e239/infoboxContent802/APE2008-Heuer.pptE2008Berlin
    - "Grant anybody, anywhere and anytime access to the (peer-reviewed)results of (publicly-funded) research"
    - Start with publications & papers archive, eventually proceed to data

  - "CERN is paving the road for a common infrastructure to allow data and resource sharing on a global scale... via the Grid."

Why LIGO? Why Now?



Why LIGO? Why Now?

- U. S.
  - NASA
    - Policy well established
    - Open data began with COBE
  - DOE
    - Congressional inquiry has started an internal re-evaluation, e.g.,
    - Genomics research (GTL) policy (April 2008):
      - "Research information obtained through public funding is a public trust. As such, this information must be publicly accessible. The GTL information-sharing policy requires that all publication related information and materials be made available in a timely manner. All Principal Investigators (PIs) within the GTL program will be required to construct and implement an Information and Data-Sharing Plan that ensures this accessibility as a component of their funded projects."
      - LHC is likely to be next.
  - NSF
    - NSB mandate linked to the Advanced LIGO authorization, but will extend beyond LIGO

Formulating An Open Data Policy for LIGO

http://www.ligo.caltech.edu/docs/ScienceDocs/M/M080072-05/M080072-05.pdf

- As a starting point, the group revisited the NSF Panel Report on the Uses of LIGO, June 1996:
  - This panel, chaired by Boyce McDaniel, produced a prescient report that foresaw pretty accurately how the Laboratory and the LIGO Scientific Collaboration would evolve;
  - Was attended by a number of current LSC members … *before there was an LSC;*
  - Anticipated the need for a phased open data policy:
    - A pre-detection and discovery phase;
      - Recognizing the need to 'protect' the data analysis process in the early phase, when the instruments are still being understood, when data analysis focuses on identifying and removing idiosyncratic instrumental artifacts that may obscure/confuse real signals
      - Recognizing that during this period experts on the instruments must be involved at every step of the science
    - An observational phase when the field of GW astronomy would begin to mature;
      - When open publication of LIGO data is a reasonable expectation

LIGO Elements Of An Open Data Policy for LIGO

- Reaffirms the McDaniel report's phased approach
  - The LSC is *de facto* an open collaboration and provides an access mechanism to LIGO data for anyone who joins the LSC.
- Initial phase: pre-detection, discovery era of LIGO
  - Progressively more astrophysically interesting upper limits leading up to the first detection
  - Additional first detections -- the 'inverse problem' era
    - Trying to fathom the physics buried in the signals
    - Continued close collaboration with selected communities through the MOU process (e.g. numerical relativity)
  - Open data proposal: release LIGO strain and ancillary data for all detections made
    - Release would be to the "broader research community"
    - Commitment: start this open data phase with the first detections, possibly as early as 2009 with Enhanced LIGO, S6.
    - Release would occur concurrently with or shortly after publication of detections
    - Possible format: consistent with NVO (National Virtual Observatory) data object
      - Snippet of calibrated time series data centered on events
      - Metadata -- extrinsic & intrinsic (derived) parameters, uncertaintie
    - Estimated impact to NSF: 1 - 2 extra FTEs, ~$0.2M

LIGO Elements Of An Open Data Policy for LIGO (ii)

- Second phase: mature era of LIGO astrophysics
  - Routine detections, populations statistics, deeper exploration of the cosmological implications, …
  - Open data proposal: continuous release of LIGO strain data, h[t], and possibly related important ancillary data;
    - Conditioned data -- artifacts either removed or tagged; known glitches tagged by time stamps, etc.
    - Commitment: start this second phase in the era of Advanced LIGO, when it is commissioned
    - Effort would require a committed core team dedicated to continuous and regular releases
    - Estimated impact to NSF: ~$4M(FY2008$)
      - ~10% of the NSF's annual investment in the LIGO GW Lab & research program
      - ~ 15+ FTEs dedicated to data products release
      - This 10% estimate is in accord with the (NASA) astrophysical laboratories which were consulted (Spitzer, SWIFT, COBE, SDSS)
      - Associated commitment to make software tools available, documented, supported, …
      - NSF acknowledges this will be new scope beyond the current level of support to LIGO GW Lab & research program
    - *Details need to be worked out*

LIGO Elements Of An Open Data Policy for LIGO (iii)

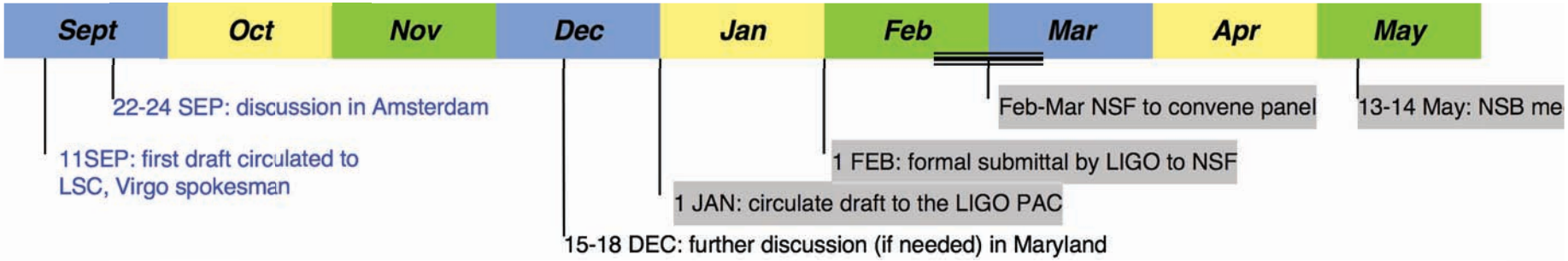
- The issues to be addressed (but we have time)
  - Iterative nature of LIGO calibrations may lead to release of imperfect data products
    - Irretrievable?
  - What is an equitable proprietary period?
    - DOE/HEP: ∞ (at present time)
    - NASA: 1 year or less;
    - LIGO: performance history suggests it takes > 2 years to complete papers from a given science run.
      - Would the NSB/NSF accept a 2 year moratorium?
      - What would we do if we were told we had 12 months to complete our "flagship analyses?"
        - Would very likely revisit our priorities and scope
  - Our MOU with GEO stipulates that GEO600 + 3 LIGO interferometers are a single network….
  - Our MOU with Virgo stipulates, "All data and their interpretations will be held strictly within the membership of the Collaboration…"

Persons/organizations who have been consulted (to date):

- Barry Barish, Member of the National Science Board, ILC Director, First LIGO Laboratory Director
- Dr. Beverly Berger, NSF Program Director for Gravitational Physics
- Edmund Bertschinger, MIT Physics Department Chair MIT
- Neil Gehrels, PI for SWIFT and Deputy Project Scientist for GLAST, NASA Goddard Space Flight Center, and member of the LIGO PAC
- George Helou, Executive Director for IPAC and Deputy Director for Spitzer Science Center, California Institute of Technology.
- Cole Miller, Professor of Astronomy, University of Maryland and member of the LIGO PAC (see below).
- Max Tegmark, MIT, Sloan Digital Sky Survey, Large Sacale Structure Analysis Team
- Rainer Weiss, Emeritus Professor of Physics, MIT, Chairman, of the COBE Science Working Group, and Founding Spokesman of the LIGO Scientific Collaboration.
- Roy Williams, Caltech, NVO Co-PI

Summary

- B. Berger(NSF) would like an endorsed proposal from the LSC and LIGO to bring before the NSF management and NSB
  - She expects that it will help make their job within the NSF easier
- We have time, but we should not be complacent
  - If we do not take the initiative to help define a solution, we may be handed one
- Timeline



- The near-term need is for NSF to report back to the NSB by April 2009 with status of what efforts have been undertaken, what provisions have been implemented.
  - ***A completely implemented solution is not needed on 1 May 2009***
  - Evidence of commitment to an open data policy and a beginnings of a roadmap to its ultimate implementation