The Phystat Physics Statistics Software Repository

M. Fischler, P. Canal, M. Paterno
Computing Division, Fermi National Accelerator Laboratory
and
J. Linnemann
Department of Physics and Astronomy, Michigan State University

March 3, 2006

Abstract

We have initiated a repository of tools, software, and technical documentation for statistical techniques used in HEP and related physics disciplines. Fermilab is to assume custodial responsibility for the operation of this Phystat repository, which will be in the nature of an open archival repository. Submissions of appropriate packages, papers, modules and code fragments from the HEP community will be made available to others in the community.

This repository is at Phystat.org. Details of the purposes and organization of the repository are presented.

Contents

Chapter 1

A Repository For Physics Statistical Code

1.1 Motivation

Currently, statistics tools are in use by individual physicists, and within collaborations. Their ultimate purpose is to make the best use of the data collected by collaborations. However, the effectiveness of these tools is in some ways limited by the lack of a straightforward mechanism for the community to share software on a wider basis, learn best practice from one another, and avoid unnecessary re-development of similar tools. Some tools (for example, event classifiers and limit calculation programs) are of general use. These codes often embody standard practices within a collaboration, recent progress of understanding within our field, or implementation of important ideas developed by statisticians or within the machine-learning communities. Other tools and programs encode hard-won expertise in handling particular subtle situations. Sharing such codes across research groups and collaborations contributes directly to the diffusion of such knowledge and indirectly to improvement of our understanding of specific data sets. Sharing this knowledge also contributes to the training of students, by facilitating the comparison of related methods.

On August 15, 2005, there was a PHYSTAT workshop at Fermilab, organized by Jim Linnemann of Michigan State and Mark Fischler of the FNAL Computing Division, to assess interest in providing an inclusive and practical mechanism for the HEP community to share statistics software, in the form of a repository for statistical software aimed at physicists. This workshop was motivated by the realization that many of the papers presented at the main PHYSTAT conferences on Statistical Problems in Particle Physics, Astrophysics, and Cosmology at Oxford[1] (2005) and SLAC[2] (2003) would benefit from a common place to cite code and place technical expositions concerning statistics techniques. There was general agreement on the need for a repository where the authors of papers in physics publications can place for citation packages containing more details about the statistical techniques underlying their analyses. A related need is for a repository which would provide, as objects of study and understanding, working codes which have been tested under realistic conditions. It was noted that such codes would also provide a point of departure for improvements, obviating in many cases the need to re-implement present ideas for lack of publicly-accessible code. Many of the participants at the PHYSTAT Conference have code module and tools which they would like to make more readily available to the physics community.

Based on consensus reached at this workshop, the idea of an archival repository aimed at physicists, with light custodial/organizational responsibilities assumed by a central institution, was presented to a wider community at the September PHYSTAT 2005 conference in Oxford. In that same time frame, preliminary discussions of a proposal to have the Fermilab Computing Division assume the role of that central institution were initiated.

As a result of these deliberations, the Fermilab Computing Division has initiated the Phystat repository. The interface for users interested in obtaining and/or submitting material to this repository is at http://phystat.org.

1.2 Did such a repository already exist?

Attendees at the August workshop represented a sizable portion of the top experts on statistical physics codes. Although some alternatives were mentioned–in particular, more than one of the attendees had at some point started a page of either links or content in this area–it was agreed that a "lone wolf" repository inherently cannot serve the functions needed. The stability and continuity implied by an institutional commitment is necessary for most of the key purposes of this repository.

That begs the question of whether the effort to create such a repository is redundant, in that an existing repository would, if publicized to the HEP community, meet the identified needs. Some possibilities discussed include:

- `arXiv.com` has no interest in holding code, and therefore is unusable for most of the purposes of the Phystat repository. Moreover, a straight copy of the arXiv mechanisms would not support many of the suggested value-added enhancements discussed.
- `SourceForge`, the `R Project` and other existing code repositories would not be suitable for code fragments, nor for expositions documenting experiments' algorithms. Moreover, the Physics Statistics code would get lost in the mass of packages in these repositories.
- `www.astrostatistics.psu.edu/statcodes` is purely a page of links, and also is aimed exclusively at astrophysics.
- `netlib.org` is mostly numerical-analysis oriented, but that in itself would not prevent using it as the repository of physics statistics code. However, quite a bit of it is dated–it is unclear how actively maintained it is.
- Repository in a Box, the technology probably used to make netlib, might be a useful tool, but it is unlikely to play well with the security constraints in today's world at Fermilab.

While it would be impractical to undertake an exhaustive search to prove that no existing repository meets the needs of the community, the overwhelming likelihood is that if such a repository existed, it would be have been familiar to at least one of the workshop attendees.

1.3 Nature of the Phystat Repository

The repository is located on an easily accessible web page phystat.org, which is managed by a content management tool (Plone[3]) which provides sophisticated navigation, search, and document management features.

The main content of the repository is a collection of "packages," each submitted by a responsible author. Each package will have been assigned a package-id, analogous to the paper-id's used in arXiv. For example, the first package accepted in March of 2006 is assigned *phystat.org/0602001*. This package-id is the recommended form of citation when citing the a Physstat package from an article.

Users can browse the contents by category, purpose, and keyword searches, obtaining lists of pertinent packages. Each package is represented by a title and one-line description, and link to a page for that contribution. On each package page, there is a description of what the code does (and/or what the technical documents are about), and a button to download the tarball(s) containing the source, documentation and build files necessary to utilize that code. All the download material is kept as part of the repository. This ensures that download links will not become broken as time passes, and retains content control against unchecked arbitrary contributor modifications.

Read access to the repository is completely open and public. Package contributors do not write directly to the repository contents. Instead, write access is be coordinated by a facilitator who will be responsible for checking content for appropriateness and for meeting the repository requirements. Thus there is human monitoring of every deposit to the publicly visible portions of this site, but (as discussed below) this filtering is a very loose moderation, rather than a "referee" process. The time between submission and public availability of an acceptable contribution is typically one or two days.

Beyond the main content of contributed packages, the Phystat repository also has convenience links to other statistics resources. And as more contributions are submitted, the repository will also contain value-added material such as platform portability and usage experience information, and/or validation and package comparisons.

1.4 Scope and organization of the Phystat repository

1.4.1 Categories of contributions

At the August Workshop and subsequent PHYSTAT Conference, several important roles for this repository were identified:

- A straightforward way to store and make available the actual software used to perform calculations for a paper (and thus defining the statistical techniques used in that paper). Journal articles could refer to this, providing unambiguous information as to how their data was refined: "we calculated the upper limit using a Bayesian technique assuming a flat prior in the cross section [17]", where reference [17] might read "H. Prosper et. al., *phystat.org/0603004*." Since the repository itself stores a down-loadable version of the code and other material, and is under the stewardship of an institution committed to continuing to make access available, such references will be stable over a very long time span.

- An associated way of storing technical documentation related to that archived software. Such documentation, though often of great potential value to others, may well be at a level of detail making it inappropriate for publication in the usual archival journals.

- A collection of useful physics-oriented stand-alone statistics utilities, and of code modules, libraries, and building blocks for assembling user-tailored statistical analyses of physics data. It was recognized that the repository should not attempt to restrict the content to some fixed or small set of languages or organizational schemes; tools and resources can be Python scripts, C++ classes, Fortran subroutines, Mathematica notebooks, or whatever form the originator found must valuable for that specific mode of usage.

- Codes implementing important new or subtle ideas, intended to be a point of departure from which others can develop superior analysis software.
- "Reference implementations" of accepted standards for statistical analysis methods applicable to physics, where such standards exist.

It was immediately realized that it is sensible to include each of these categories in the repository. This is reflected in the set of categories appearing in the navigation box at http://phystat.org:

- Libraries & Modules
- Code Fragments
- Technical Documents
- Toolsets

In a repository with this degree of variety in content, it is natural to take an inclusive approach in its acceptance policy. Therefore, the Phystat repository defines some basic expectations concerning what a submitting author should provide (such as a synopsis identifying what the submission is, and the actual code–if any–of the submission), but keeps the requirements to a minimum, in order to encourage physicists to submit content. The decision that the repository should be as inclusive as possible has a consequence regarding organizational responsibility: In order to maintain the repository's practical usefulness, and prevent it from evolving into an incomprehensible jungle of wildly varying tools, at least some light degree of categorization, organization, and content road-mapping must be provided by the repository custodians.

1.4.2 Purposes of the contributions

The areas of concern are codes and techniques involving data fitting, statistical limit setting, data/event categorization, hypothesis testing, pseudo-random number and distribution generation, and graphical display of results of statistical analysis. Each package should address one of these purposes; the process of submitting a package includes selecting a relevant purpose from a list which includes the above areas.

Material in the Phystat repository should also be relevant to physics applications. Although it is anticipated that much of the early contributions will be from the HEP, particle-astrophysics, and Lattice QCD communities, the repository will be open to material relevant to other physics applications as well.

The above list of purposes is intended to define the scope of the Phystat repository, but it probably is not complete yet. If the physicists feel that there are benefits to including other areas, for example using the repository as a way to share tracking algorithms, such input will be appreciated. Since the repository is set up for easy browsing along various views, expansion of the set of package purposes wouldn't necessarily introduce much clutter for the user who is looking for statistics packages.

Most items in the repository will consist of contributed code and documentation of that code and/or the statistical techniques or algorithms embodied. We take a broad definition of "code", which also includes Root macros, scripts, macros (for example, Mathematica or Excel macros), and even useful database table formulations and SQL techniques.

This is a *software* repository, so material which is not at all associated with code should not in general be submitted. However, it is perfectly appropriate to submit (for the Technical Documents section of the repository) papers which:

- explain the motivation or mathematical or technical details of a particular code or related set of codes in use in the community
- consist of consist of explanatory material concerning algorithms or techniques utilized by multiple items in the repository (and thus would not be appropriate to associate with any specific one of those packages)
- explain the motivation or mathematical or technical details of a particular code or related set of codes in use in the community
- provide for archival purposes data sets, programs, and scripts sufficient to reproduce figures and results appearing in a publication
- provide mathematical justification or background explanation for techniques in general use, as applied to physics applications
- propose or define statistical approaches and schemes to address various problems, intended for eventual migration (by the author or otherwise) into actual code

The strategy for this repository is:

1. We should be as inclusive as possible: No restrictions based on which platforms or languages a package works with; no acceptance/refereeing wrestling; the broadest possible acceptance of licensing approaches.
2. Institutional responsibility is a key point: To ensure that archived material will remain available over a long time span. Assigned package numbers will be suitable for use as citations, without concern that they will become invalid.
3. We should not be too ambitious: The repository material will come from the community, not from some core of Phystat maintainers/developers.
4. Universal download access is important: Browsing and searching must not require passwords, certificates, or presence on some privileged network.
5. Contributors cannot remain anonymous, but anybody with a suitable package can (by supplying a valid return address) become a contributor.

Chapter 2

Using the Phystat Repository

The first step in using the Phystat repository is to go to its URL, at phystat.org. (Depending on your browser and the security nature of the networks involved, you may be asked to choose among some certificates, but no authentication certificate is needed. Simply select the "cancel" option on the question about certificates, and you are in for the purposes of browsing and downloading.)

On this page, you will find:

- A top tab-bar, including the important "package search" tab.
- A navigation box on the left. You can use this to brows just Libraries and Modules or just Code Fragments or just Toolsets.
- A login area (and a "join" link in the upper right). Users intending to browse and/or download material need not log in. You would have to join and login in order to submit a package for others to use.
- The central page, with
  - A "How To" area with instructions for searching, submitting and updating a package, and commenting on packages.
  - Information about the repository itself.
  - Links to all the PHYSTAT conference pages.
  - Convenience links to other statistics resources.

2.1 Finding Packages

Two modes of user searches for packages are anticipated. The user wishing to look up a package cited in some article will have at hand the assigned package-id found in the citation. Entering that package-id will immediately move the user to the package page for the selected package.

Other users will wish to examine the contents of the repository, looking for material which is for relevant to their needs. The package search page allows the user to specify any combination of package type, programming languages, purposes, keywords, and authors. If you know the title desired, you can search by title. The search engine also accepts description words, and provides general text search capabilities.

Any such search returns a list of pertinent packages. (These package lists are created dynamically at search time; thus if a new package is contributed to the repository and accepted, the repository maintainers need do nothing extra to ensure that the new package will show up on subsequent searches in all appropriate views.) The list includes the short descriptions provided by the submitting authors. Selecting one of the listed packages moves to the package page. There you will find a more detailed description of what the code does (or what the paper is about), and a button to download the tarball(s) containing the source, documentation and build files necessary to utilize that code or paper.

2.2 Submitting a Contribution

It is recommended that the potential contributing author click the "How To" link for submitting a package. The instructions there will guide the process, including the steps to join the site if there is no existing login available, and the steps to enter the package description and download the materials.

The package submitted should be suitable for the repository:

1. The material must be relevant to statistics, data fitting, or random numbers and must pertain to or be useful for physics applications.

2. The contribution must include code (within the inclusive definition presented in §1.4.2) or explanatory material concerning algorithms or techniques utilized by items in the repository.

The author should prepare the needed materials:

1. A submission should include identification of author(s), and an e-mail contact address.

2. The contributing author must choose a title and a short (one line) descriptive phrase, suitable for browsing.

3. The contributing author must provide a brief description suitable for a perspective user to read to decide whether to download and explore using this the code or material.

4. The contributing author must provide a tarball containing all the code, build files, documentation, data sets, and other material composing this contribution.

5. The contributing author will have to agree to give the repository permission to distribute the contributed code and hold the repository and Fermilab blameless for actions by users who download that code. (The author is free to insert copyright and/or license restrictions in the material, as long as the repository is given permission to distribute, and it is understood that all enformement activities are the responsibility of the author.)

In accordance with the "come as you are" philosophy, the Phystat repository imposes no further requirements. The idea is to make it as easy as possible to submit a package, without having to fill in material which in the opinion of some set of arbiters must be present for a package to be "good." Additional suggested (but not required) information will include a `README` file to describe how to get started in using the code, information about dependencies and how to obtain the needed additional software, keywords to make user searches more fruitful, experiment of origin information, and the platform(s) and circumstances on which this code has been run.

With this material in hand, the steps to submit a package take about three minutes. Since the repository content is loosely moderated, you can expect the submitted package to become publicly visible in a day or two. The moderators are not trying to be judges of quality; any appropriate package within the scope of Phystat will be accepted.

A separate but similar mechanism applies for submitting established for revisions and updates of contributed items. Access to past-version tarballs will be retained.

Chapter 3

Activities and Policies

3.1 Steering Committee

Decisions about repository policy would be made by a small steering committee consisting of repository supporters and external initiators, physics statisticians, Fermilab CD (and/or other) management, and representatives of various experiments and large physics software projects.

The steering committee will "meet" primarily by email, and should be in place in time to guide the priorities of "value-added" activities. Current members include Jim Linnemann, Louis Lyons, Mark Fischler, Harrison Prosper, Glen Cowan, and Kyle Cranmer.

3.2 Potential value-added activities

A longer term vision of the repository goes beyond passively archiving code, useful though that is. Phase II of the project will consist of steps to enhance the value of the repository, both to users accessing contributions, and to contributors benefiting from acknowledgment that their contributed code is useful to others.

Some of these steps would naturally be done by the repository moderators. Other activities depend on participation by outside physicists, either organized by the repository or undertaken as individual efforts. These activities are all potential, depending on community desires and time available.

The guidance of the steering committee will be important in the priorities assigned to the following (and potentially other) value-added activities.

3.2.1 Classification and validation activities

- Classification of the submissions to distinguish archival entries from actively maintained packages.

- Presentation of an easy way to attach community comments, and organization of this user feedback.

- Basic validation and certification, to identify tools and modules which are more suitable to out-of-the-box use and those which are more points-of-departure for further code development.

- Organization of community comparisons among packages.
- Further dimensions of package classification, including for example by platform availability.

3.2.2 Extending the scope and contents

- Presentation of a “code wanted” list, where the community can express needs for specific capabilities.
- Actively looking for interesting and relevant software produced by the statistical software community, and providing web interfaces or language translation wrappers to support use by the physics community.
- Providing and maintaining a list of statistics software (beyond the contributions to the Phystat repository) and key papers.
- Identifying missing functionality and providing code to fill the gaps.

3.2.3 Improving capabilities

- Soliciting and supporting extensions of existing code (justified by making possible a broader user-ship than a single experiment).
- Integrating related packages.
- Provision of design expertise to create standards to make packages more readily usable.
- Improvements on contributed code, to enhance either portability, efficiency, or maintainability.

Chapter 4

Summary

The Phystat repository for statistics-related physics code is operating, but it is brand new. The contents have been growing by about one submitted package per week. The next important step is for physicists to make use of the repository. Browse for packages you may be able to use. Browse to see how various experiments tackled your statistics issues. Use the repository to download versions of important packages.

The repository will be what the community makes it. Validation and endorsement comments are very useful, even before any formalized mechanisms are in place. And it is essential that users report any problems and make suggestions for improvements. In particular, comments about the mechanics of using the repository are welcome.

The HEP community can make physstat.org valuable. Submit packages to be disseminated. Submit code fragments defining how your analysis did statistics. (These can then be cited by Phystat package-id number, much like a paper in arXiv is cited.) Submit technical documents explaining choices of statistical approaches. There is a large backlog of code and tools potentially valuable to the HEP community.

Bibliography

[1] PHYSTAT O5 - Statistical Problems in Particle Physics, Astrophysics, and Cosmology. www.physics.ox.ac.uk/phystat05/ Oxford, UK, September 12-15, 2005.

[2] PHYSTAT 2005 - Statistical Problems in Particle Physics, Astrophysics, and Cosmology. www.slac.stanford.edu/econf/C030908/

[3] The Plone Open Source Content Management System. plone.org