HEP Software & Computing Knowledge Base

hepsoftware.org

Torre Wenaus, BNL
CHEP 2016, San Francisco
October 13, 2016
Why a HEP S&C knowledge base?

Number one on the original HEP Software Foundation (HSF) task list: “provide a system for facilitating information exchange”

Proposition accepted at HSF’s formative workshop at SLAC in 2015: HEP S&C should have a grass roots information gathering and exchange hub

- Facilitate collaboration by increasing awareness of resources, projects, activities
- Provide a place to look for solutions, learn from the choices made by others
- Promote awareness of software and tools from outside our community, e.g. open source

A resource built and maintained by the community

- Must overcome the Catch 22: It’s only useful if it’s populated, and people will only help populate it if it’s useful
- Entice people to contribute by making it fun (or at least pleasant) to do so
hepsoftware.org development

- Key design precepts
  - Entity-relation model: relations are as important as entries, much of the useful knowledge is in the interconnections (e.g. who uses what)
  - Make it fast, highly navigable, easy transition between browsing and editing
  - Users are also creators

- Goals reached in the third prototype generation
  - 2014: Django, MySQL, xml data, based on ATLAS PanDA monitor. Clunky
  - 2015: Drupal + extensions, part of HSF website. Slow and clunky
  - 2016: Current version based on ATLAS ‘data knowledge base’ prototyping
  - 2017: No plans for a fourth, just clean up the code

- Data content carried forward and preserved generation to generation

- Implementation: a javascript app in the browser served by a node.js server
  - MySQL data repository
  - REDIS memory store for server side sessions
  - CouchDB/PocketDB for server/client json data channel
  - Works beautifully

- In github private repo, move to public when purged of secrets and housecleaned
hepsoftware.org operation

- Two knowledge base usage modes:
  - Data subset at the client with queries to server for the rest
  - or, Fully client-resident data
  - Currently uses the latter; data volume is small
    - Content is all human typed text
- Non-negligible initial load time pays off in responsiveness
  - plus it works offline
  - 5 sec load time in my hotel room this morning
- Has operated very stably for close to a year
  - Server in Amazon EC2, provided by BNL
  - Backed up every 2 hours to EBS and S3, daily backups kept
  - Full revertible version histories for entries kept

Content today: 603 entries, 963 relations created by 24 people
The remainder of the talk is a lightning tour until my time runs out…

…what I don’t get to is supplementary
Getting help

Click the info button for a guide, which is itself an entry in the KB

All KB content is in markdown

Navigable table of contents for large entries

Licensed, with its content, under Creative Commons Attribution-ShareAlike
Entries have attributes

- Contact people
- Website, twiki, wikipedia
- Email fora, blog
- Social media
- Events, event series
- Training
- WikiToLearn
- Reference links
- Documentation
- Presentations
- Downloads
- Repository
- Issue tracker
- Testimonials
- Jobs
- ...

**Attributes**

- **Url** PanDA twiki
  - https://twiki.cern.ch/twiki/bin/view/PanDA/PanDA
- **Url** PanDA overview
- **Url** BigPanDA project
- **Url** Publications, presentations, news
- **Github** Github repository
  - https://github.com/PanDAWMS
- **Url** PanDA monitor
  - http://bigpanda.cern.ch

**Contact**

Kaushik De (UTA), Alexei Klimentov (BNL), Torre Wenaus (BNL)

Add an attribute of type:

- Website
- Contact
- Wiki
- Documentation
- Reference
- Presentation
- Repository
- GitHub
- CERN GitLab
- Git
- Bitbucket
- Issue tracker
- Download
- License
- Forum
- Blog
- Event
- Event series
- Training
- WikiToLearn
- Jobs
- Wikipedia
- YouTube
- Twitter
- Reddit
- Content credits
- Image
- Logo
- Testimonial
- Other
Entries have relations

Key to the entity-relationship organization of the knowledge base

Ascribe relations between entries of contextually appropriate types

Builds up the *knowledge* part of the KB

Convey connections, enable discovery

Share the wisdom of your experiment or project in what you use

There are many more relations in the database than there are entries
Adding and editing material

Live preview
In-app, fast, responsive
Convenient autocomplete selection from long lists like software categories
Cached at client until save, no “frack! my form is gone” when your connection dies
Markdown based content
Multiple active edits OK

Lightning Tour
Software project entries are organized by category
An entry can belong to one or many categories
Add your favorite software projects, and/or improve their entries!
Entries related to the selected (red) entry are indicated (blue)

Category entry gathers info on the subject
(With levity courtesy of xkcd)
Linked articles

- Software quality
  - Software sustainability
  - Software licenses
  - Statistical tools
  - Supercomputing
  - Training software
  - Trigger/DAQ
  - User
  - Environment Management
  - Virtualization and clouds
  - Web
  - Web app frameworks
  - Web based tools and services

- Software quality category

- Parent
  - Software engineering

- Related
  - Developing maintainable software | Software Sustainability Institute
  - The Beauty of Code - Paris Review
  - Writing readable source code | Software Sustainability Institute

Lightning Tour
Entries for mature and embryonic projects
Both are appropriate
Reference for the well established
Means of discovery for the new
References to doc, web, tools, repo, license, forum, events, social media, ...

Associated institutes, groups, software
Describe your experiment and the software it uses

Valuable resource for new projects surveying who uses what
High intensity, neutrino related

- Art event processing framework
- artdaq
- BES III
- CAPTAIN
- COMPASS
- Daya Bay
- DD4hep
- DUNE
- Fabric For Frontier Experiments (FIFE)
- fast
- Geant4
- GENIE
- HARP (PS214) - The Hadron Production Experiment at the PS
- Heavy photon search
- KOTO
- LArAT
- LArSoft

Lightning Tour
Describe your favorite organization, relate it to software projects, facilities, training programs...
A community of scientists, researchers, and experts in high throughput computing around the world. The OSG represents US LHC and the US LHC cyberinfrastructure.

2015 proved to be an exciting year for the Open Science Grid Operations Center. Milestone of one billion CPU hours delivered (last 12 months) was reached in October. 100 million hours per month was exceeded for the first time in November and the target was reached in March.

Production services for operation of access to heterogeneous research facilities and over 100 universities:
- US contribution to the successful World Wide LHC Computing Grid.
- 2012-2016 28FTEs supported by DOE and NSF. Used continuously since 2009. Opportunistic sharing of available resources across a mix of sciences.
- Past 12 months: over 2.0M CPU hours and 1 Petabytes transferred per day.
- Expert consulting for users, developers, and resource administrators.

Software Sustainability Institute

The research community is ever more reliant on software, much of which is developed within the community. The Software Sustainability Institute supports better research by helping researchers to build and use better software.

Associated with

- Depsy

Related

- Software sustainability
- Training software
- Approaches to software sustainability | Software Sustainability Institute
- Choosing a repository for your software project | Software Sustainability Institute
- Choosing an open-source licence | Software Sustainability Institute
- Developing maintainable software | Software Sustainability Institute
- Supporting open-source software | Software Sustainability Institute
- The researcher programmer, a new species? | Machine Doing
- Writing readable source code | Software Sustainability Institute
Objective: draw CHEP content into the cross-referenced knowledge base, have it accessible, visible, discoverable, integrated

A start made with CHEP 2015, 2013 but only to the session & proceedings track level

Add your own talks & proceedings as links and associate them with their session, sw category etc
Events can be entered as attributes on an entry, or as entries themselves

New: events are imported from the new & popular HSF community google calendar

Event series entries can be associated with individual events

<table>
<thead>
<tr>
<th>Event</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMS Offline &amp; Computing Week, CERN</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>XrootD Workshop, Tokyo, Japan</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>ALICE Offline Week, CERN</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>IEEE/NSS MIC Conference, Strasbourg, France</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>Geant4 Review</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>ESC16 School, Bertinoro, 47032 Bertinoro, Province of Forli-Cesena, Italy</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>ATLAS Week, CERN</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>HEPiX Fall 2016 Workshop, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720, USA</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>LCLS II Collaboration meeting and Review</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>CHEP 2016, San Francisco, CA, USA</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>LHCb Analysis and Software Week</td>
<td>from HSF calendar</td>
</tr>
<tr>
<td>CHEP 2016, San Francisco</td>
<td>from CHEP 2016, San Francisco</td>
</tr>
<tr>
<td>4th Workshop on Sustainable Software for Science</td>
<td>from 4th Workshop on Sustainable Software for Science</td>
</tr>
<tr>
<td>Higgs Hunting 2016, Orsay</td>
<td></td>
</tr>
<tr>
<td>CERN Openlab/Intel hands-on workshop on code optimization</td>
<td>This workshop — organised in collaboration with Intel — is a great opportunity to improve your application’s performance and ask questions to Intel Software Architects!</td>
</tr>
<tr>
<td>2016-04-05</td>
<td></td>
</tr>
</tbody>
</table>

Lightning Tour
Institutes

Universities and labs, with relations to the software, S&C groups, services, facilities, experiments etc. that are associated with them

Brookhaven National Laboratory (BNL)

Brookhaven advances fundamental research in nuclear and particle physics to gain a deeper understanding of matter, energy, space, and time; apply photon sciences and nanomaterials research to energy challenges of critical importance to the nation; and perform cross-disciplinary research on climate change, sustainable energy, and Earth's ecosystems.

Associated with:
- BNL Electronic Detector Group
- BNL Physics Applications Software Group (PAS)
- BNL RHIC ATLAS Computing Facility (RACF)
- HEP Software Foundation (HSF)
- hepsoftware.org
- STAR

Lightning Tour

HEP S&C KB  T. Wenaus  CHEP 2016
Describe your group as an organization, relate it to your software projects

**BNL Physics Applications Software Group (PAS)**

**Organizations**

The Physics Applications Software (PAS) group in BNL’s Physics Department develops software in support of BNL’s HEP program. The largest PAS effort goes to the ATLAS project, but other members also contribute to the DUNE neutrino physics collaboration and the LSND project. The group also maintains software for the LHC Computing Grid (LCG) and the CERN Physics Data System (CDS).

**Website**

Contact Alexei Klimentov, PAS Group Leader

**Associated with**

- ATLAS Event Service
- BNL’s PAS group is a lead contributor to the Event Selection System
- Brookhaven National Laboratory (BNL)
- Physics Applications Software (PAS)
- hepsoftware.org
- PAS developed and hosts hepsoftware.org
- NICOS nightly build system
- NICOS was developed by BNL’s PAS Group
- PanDA
- PAS is a lead contributor to PanDA
- CERN
- CernVM
- Concurrency Forum
- GENSER - Generator Service Project
- HepMC3
- Jupyter@CERN
- LHC Computing Grid (LCG)
- ROOT
- Software Technology Forum

**Reporting they are used by CERN EP-SFT**

- CernVM-FS / cvmfs
Resources

Computing facilities, online services, publication series, online training series and materials, career resources, funding agencies & opportunities

BNL RHIC ATLAS Computing Facility (RACF)

Resources/Facilities
The BNL RHIC ATLAS Computing Facility (RACF) at Brookhaven National Laboratory provides computing services for the experiments at the Relativistic Heavy Ion Collider (RHIC) at BNL, the US-based collaborators in the ATLAS experiment at the Large Hadron Collider at CERN, and the collaborators in the Large Synoptic Survey Telescope project.

In ATLAS, RACF serves as the ATLAS Tier 1 Facility in the US, the largest such facility in ATLAS providing close to a quarter of ATLAS computing power. It serves as a hub for the US ATLAS Integrated Facility which includes five Tier 2s spanning 10 institutes.

- Website
- Contact
- Guided tour

Associated with
- AGLT2 - ATLAS Great Lakes Tier2: AGLT2 as a Tier 2 is associated with the BNL Tier 1
- ATLAS: ATLAS Tier 1 Center
- Brookhaven National Laboratory (BNL): Located at BNL
- Ceph: RACF provides an object store backed by Ceph
- LSST Dark Energy Science Collaboration (DESC): A computing facility for LSST
- NX at the RACF
- Open Science Grid (OSG): RACF interfaces to the OSG
- PHENIX: Principal computing facility for PHENIX
- STAR: Principal computing facility for STAR

Created 2015-11-05 by wenaus, updated 2015-11-08 13:00 by wenaus

Lightning Tour
Describe your Tier 1,2,3 and its projects

AGLT2 - ATLAS Great Lakes Tier2

AGLT2 - ATLAS Great Lakes Tier 2 provides computing and storage capacity for US ATLAS physicists running ATLAS simulations and data analysis. We currently provide more than 4000 total job slots (6600 cores) and 3.7 Petabytes of storage capacity interfaced to the Open Science Grid. Our job scheduling system is Condor and we utilize dCache, Lustre, AFS, and NFS as storage systems.

AGLT2 is a federated Tier 2, comprising facilities at U Michigan and Michigan State, operating as a single unified facility.

AGLT2 is one of the top ATLAS Tier 2 sites worldwide in terms of CPU hours provided. In any given 24 hour period we average more than 8000 completed ATLAS jobs. We transfer an average of 1-2TB of data per day to and from other sites.

AGLT2 is also an ATLAS Muon Spectrometer Calibration Center.

Networking

AGLT2 is a leader in advancing networking, including advancing perfSONAR as a basis for uniform network performance monitoring across LHC computing resources.

PunDIT

AGLT2 is a collaborator on the PuNDIT project which will integrate and enhance several software tools to provide an infrastructure for identifying, diagnosing and localizing network problems. In particular, the core of PuNDIT is the Pythia tool that uses perfSONAR data to detect, identify and locate network performance problems.

OSiRIS

AGLT2 is currently developing OSiRIS, a distributed, multi-institutional storage infrastructure that will allow researchers at any of our campuses to read, write, manage and share their data directly from their computing facility locations.

Our goal is to provide transparent, high-performance access to the same storage infrastructure from well-connected locations on any of our campuses. We intend to enable this via a combination of network discovery, monitoring and management tools and through the creative use of CEPH features.

By providing a single data infrastructure that supports computational access on the data "in-place", we can meet many of the data-intensive and collaboration challenges faced by our research communities and enable these communities to easily undertake research collaborations beyond the border of their own Universities.

Associated with

- BNL RHIC ATLAS Computing Facility (RACF)  AGLT2 as a Tier 2 is associated with the BNL Tier 1
- Open Science Grid (OSG)  AGLT2 interfaces to the OSG

Website  Contact  Wiki  OSiRIS  PuNDIT

Lightning Tour
Licenses

Software category gathering licensing information

License is an attribute on software projects

The license tag shows what products have declared a license
The Helmholtz Alliance “Physics at the Terascale” bundles German activities in the field of high-energy collider physics. It is a network comprising all German research institutes working on LHC experiments, a future linear collider or the related phenomenology.

- **Website**
  - Advanced programming concepts 2016  2016-03-07
  - Terascale statistics school 2015  2015-03-23
  - HistFitter tutorial  2015-03-30
  - Computer Algebra and Particle Physics - CAPP 2015  2015-03-30
  - Monte Carlo School 2015  2015-04-13
  - Terascale C++ School 2015  2015-06-08
  - Fast Monte Carlo Workshop  2014-01-14
  - Terascale Monte Carlo School 2014  2014-03-10
  - Terascale C++ School 2014  2014-06-16
  - GPUs in HEP  2014-09-10
  - Computer Algebra and Particle Physics - CAPP 2013  2013-03-18
  - Introductory Statistics School 2013  2013-03-18
  - Monte Carlo Methods in Advanced Statistics Applications and Terascale Statistics School 2016  2016-02-15

Created 2016-01-23 by wenaus

- **Training series**
- **Schools**
- **Commercial online training**
- **Software product training**
- **Community resources**
- **Material gathering points**

**Lightning Tour**
Depsy: career resource example

Depsy helps build the software-intensive science of the future by promoting credit for software as a fundamental building block of science.

Depsy text-mines papers to find fulltext mentions of software they use, revealing impacts invisible to citation indexes like Google Scholar.

Citation is just part of the story—Depsy analyzes code from over half a million GitHub repositories to find how packages are reused by other software projects.

Depsy assigns fractional credit to contributors based on designated authorship, number of commits, and repo ownership—supporting a fairer, more software-native reward system.

Depsy currently works for the 11,223 Python and R research software packages available on PyPI and CRAN.

Associated with

<> Impactstory  Depsy is built by Impactstory

Science Code Manifesto
Software Sustainability Institute

Related

The unsung heroes of scientific software: Nature News & Comment
The researcher programmer, a new species? | Machine Doing
Recent addition not much used yet but powerful

Full flexibility to invent classifications
Links

You can add *anything* as a link (as long as it’s a link)

When you read a cool software design article, add it as a link

eg to relate a CHEP track’s proceedings to its software category, add it as a link
Instant results
Fast-scan navigation among results; click to select
Search by title and/or content
Empty search shows all entries, sorted by alpha or modification time
Accounts, authentication

- No internal user/pass database
- Authentication based on email address
- Obtained from your choice of third party
  - GitHub, Google, Dropbox, Amazon, ...
  - As you often see these days on websites
- Missing is CERN, they have a new OAuth2 service
- REDIS based server-side session management
- Once authenticated, you have an identity in the KB, associating you with content you create and modify, testimonials you add, preferences you set
  - People not supported as entries properly yet
  - You should be able to relate people to other entries (software, experiments, institutes, …) but can’t yet
- Perhaps more personalization in the future
  - Experiment association, enabling experiment-specific and -internal content
  - Notifications of new content… RSS? email?
Versions, visibility

Previous versions are archived and can be viewed and restored as the published version.

Nothing can be deleted by the user, you can hide items.

Versions

Previous versions are accessible here. They can be loaded as the edited version (so long as doing so would not overwrite active edits), and then saved, in order to revert to previous versions.

Archive 2014-10-20/wenaus, updated 2015-11-11/wenaus
Archive 2014-10-20/wenaus, updated 2015-11-11/wenaus
Archive 2014-10-20/wenaus, updated 2015-11-05/wenaus
Archive 2014-10-20/wenaus, updated 2015-11-04/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-27/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-27/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-26/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-25/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-25/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-15/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-15/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-07/wenaus
Archive 2014-10-20/wenaus, updated 2015-10-07/wenaus
Archive 2014-10-20/wenaus, updated 2014-10-20/wenaus
Next steps

Add and improve content! An on-and-off hobby of mine but it needs you

- New entries, testimonials on software, dressing entries with tags and relations, adding that cool article on online training you found, adding your training event, ...

In terms of development, best effort and driven by interest

- Ease of use, aesthetics, performance
- Clean up the code and open it up in github
- People as full fledged entries with associations
- People-experiment association to enable experiment-specific (including private) material
Summary

The HEP S&C knowledge base at hepsoftware.org is there for you to try, as a user and as a content contributor.

Stable, functional, zippy, even kind of fun to use and add stuff when you get in the swing of it.

Benefits from a couple rounds of prototype cycles.

Smart javascript client locally hosting the data makes for a very responsive dynamic browser app.

Node.py + MySQL based back end in EC2 has been solid & stable.

What it needs most now is content!