Analysis Preservation for ATLAS

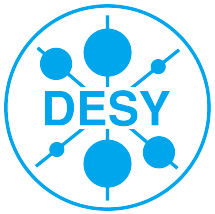
Conclusions of the Analysis Preservation Panel
https://cds.cern.ch/record/2242594/

David South (DESY)

ATLAS Data Characterization
and Curation Meeting

August 24th 2017

HELMHOLTZ
ASSOCIATION



ATLAS AP panel discussions arrived at two strategies

One: To capture in as much detail as possible a description of the analysis including the triggers, object definition, overlap removal, event selection, background Monte Carlo samples, data driven backgrounds and so on. This approach will provide the possibility to recreate the analysis, and thus ***reproduce*** and/or ***re-use*** the analysis. The analysis level data also comes here

Two: Encapsulating the actual code and workflow that was used for the analysis organised so that it can be re-run exactly as before, to faithfully ***reproduce*** the analysis and/or provide opportunities for ***re-use*** and reinterpretation

- The two strategies are nevertheless complementary to each other
  - Strategy One has perhaps more emphasis on **reproducibility**
  - Strategy Two is more about a practical solution to **re-use** an analysis
- Both strategies can be implemented by ATLAS as part of the CERN Analysis Preservation (CAP) framework

CERN Analysis Preservation framework and working group

- In summer 2014 a first meeting took place to gather feedback on the development of the Data Analysis Preservation Framework (DAPF)
- This became the CERN Analysis Preservation (CAP) Framework in 2015
  https://github.com/cernanalysispreservation/analysispreservation.cern.ch/wiki/Overview
  - Group based in CERN-IT, with participation from all LHC experiments
  - Links to DPHEP, DASPOS, RECAST and others
  - Workshop in February:
    https://github.com/cernanalysispreservation/analysispreservation.cern.ch/wiki/Joint-CERN-Analysis-Preservation-DASPOS-RECAST-workshop
- The CERN Analysis Preservation efforts focus on three pillars:
  - **Describe** aims at capturing the structured knowledge information about data analysis processes
  - **Capture** aims at capturing the computing environment, the software code, the datasets, the configuration and other information assets used in data analyses
  - **Reuse** aims at re-instantiating preserved analyses on a containerised computing cloud for the purposes of re-validation and re-interpretation

CERN Analysis Preservation framework and working group

- In summer 2014 a first meeting took place to gather feedback on the development of the Data Analysis Preservation Framework (DAPF)
- This became the CERN Analysis Preservation (CAP) Framework in 2015
  https://github.com/cernanalysispreservation/analysispreservation.cern.ch/wiki/Overview
  - Group based in CERN-IT, with participation from all LHC experiments
  - Links to DPHEP, DASPOS, RECAST and others
  - Workshop in February:
    https://github.com/cernanalysispreservation/analysispreservation.cern.ch/wiki/Joint-CERN-Analysis-Preservation-DASPOS-RECAST-workshop
- The CERN Analysis Preservation efforts focus on three pillars:

ATLAS strategy one

  - **Describe** aims at capturing the structured knowledge information about data analysis processes
  - **Capture** aims at capturing the computing environment, the software code, the datasets, the configuration and other information assets used in data analyses
  - **Reuse** aims at re-instantiating preserved analyses on a containerised computing cloud for the purposes of re-validation and re-interpretation

CERN Analysis Preservation framework and working group

- In summer 2014 a first meeting took place to gather feedback on the development of the Data Analysis Preservation Framework (DAPF)
- This became the CERN Analysis Preservation (CAP) Framework in 2015
  https://github.com/cernanalysispreservation/analysispreservation.cern.ch/wiki/Overview
  - Group based in CERN-IT, with participation from all LHC experiments
  - Links to DPHEP, DASPOS, RECAST and others
  - Workshop in February:
    https://github.com/cernanalysispreservation/analysispreservation.cern.ch/wiki/Joint-CERN-Analysis-Preservation-DASPOS-RECAST-workshop
- The CERN Analysis Preservation efforts focus on three pillars:
  - **Describe** aims at capturing the structured knowledge information about data analysis processes
  - **Capture** aims at capturing the computing environment, the software code, the datasets, the configuration and other information assets used in data analyses
  - **Reuse** aims at re-instantiating preserved analyses on a containerised computing cloud for the purposes of re-validation and re-interpretation

ATLAS strategy two

Recommendations with conclusions and next steps

1. *The panel recommends ATLAS pursue a course of analysis preservation to allow analysis reproducibility and reuse of data, tools and concepts. This should be done in close coordination with the CAP group and the other stakeholders, primarily made up of the other LHC experiments*

- A significant number of **use cases** have been established in favour of analysis preservation, where both reproducibility and reusability of ATLAS analyses by the collaboration are applicable

- The two identified complementary strategies now require further, broader investigation and suitable candidate analyses should be identified by Physics Coordination from each of the physics working groups:
  BPHYS, EXOT, HIGG, STDM, SUSY and **TOPQ**

- The development and implementation of the projects proposed should proceed within the newly established *ATLAS Data Characterization and Curation* (DCC) working group

Recommendations with conclusions and next steps

2. *The CAP Portal provides an ideal tool for capturing all information about an ATLAS analysis and this should be evaluated to the fullest extent*
3. *The automated upload of metadata, primarily but likely not exclusively from Glance will require further development from the ATLAS side, working closely with CAP. This is a critical requirement to the success of the project*
4. *Furthermore, once available, the upload of data associated to an analysis to the CAP storage should also be investigated*

> The current prototype of the ATLAS CAP interface is very basic and can only be expanded upon once a **wider metadata appraisal** has been completed by the physics working groups

> The LHCb and CMS interfaces are more mature and ATLAS should collaborate with them via the CAP team to benefit from their experiences

Recommendations with conclusions and next steps

2. *The CAP Portal provides an ideal tool for capturing all information about an ATLAS analysis and this should be evaluated to the fullest extent*
3. *The automated upload of metadata, primarily but likely not exclusively from Glance will require further development from the ATLAS side, working closely with CAP. This is a critical requirement to the success of the project*
4. *Furthermore, once available, the upload of data associated to an analysis to the CAP storage should also be investigated*

- Six months of funding is available from CERN RCS-SIS for a **dedicated ATLAS position working with the CAP team** to implement the ATLAS interface. This person would be working closely with the Glance development team (would be beneficial if this is matched by ATLAS, to make a 1y position)
- **This will cover**: the development of the CAP back-end required to talk to ATLAS databases such as Glance, commissioning of the “one shot” analysis and workflow uploading functionality and the integration of the dedicated storage space for the data associated to each preserved

Recommendations with conclusions and next steps

5. *The successful integration of analysis workflows into the CAP and their retrieval should be further explored and expanded*

- The preparation of ATLAS analysis workflows for capture, including all relevant software, tools, steps as well as the working environment, should be investigated to cover **a more generic analysis base**

- Whilst the focus so far has been on the reinterpretation of SUSY analyses, establishing if such workflow definitions can be defined and provided by all types of analysis performed by ATLAS demands **a more detailed evaluation by the physics working groups**

6. *Once in place and successfully validated, the upload of an ATLAS analysis to the CAP, including metadata, data and the analysis workflow and environment should form part of the publication sign off procedure, if shown to be technically feasible*

Back Up

Analysis preservation: Not data preservation, not open data

- ATLAS is involved in and has policy on data preservation and open data
  - ATLAS Data Preservation Policy, which outlines the **general principles of data preservation**: the data themselves, data formats and reproducibility of physics results https://indico.cern.ch/event/211843/contribution/12/material/0/0.pdf
  - The requirements for preserving ATLAS data **for future use by the ATLAS Collaboration**: ATL-SOFT-INT-2014-001, https://cds.cern.ch/record/1697900?ln=en
  - ATLAS policy on **data access** rules, based on the established DPHEP preservation levels: https://indico.cern.ch/event/286440/contribution/7/material/0/0.pdf
  - The ATLAS **open data** project: http://opendata.atlas.cern

- Analysis Preservation is somewhat different, where the focus is on the analyses performed by the experiment, rather than the data themselves
  - It is for the benefit of the collaboration, and is not related to the public release of data

- Some necessary re-definitions agreed upon by the panel:
  - ***Reproducibility*** is defined as repeating an analysis of a given data set using the original procedures, software and tools: **"to reproduce what was done before"**
  - ***Reusability*** is defined as repeating an analysis using new data and/or new versions of old data, potentially with new software and/or tools: **"to re-do an analysis, or to re-use it for new purpose"**

Key points from the mandate

- *The panel should discuss the validity of [given] use cases and identify others*
- *The panel should consider whether preservation should mean "reproducibility" or ["reusability"] or both*
- *The panel should make recommendations about how analyses could be preserved in ATLAS, and comment on whether any tools ... are adequate for this purpose*
- *The panel may lay out several options and describe the benefits and the effort needed for each*
- *The panel should solicit input from both within ATLAS and outside ATLAS*

Members of the panel

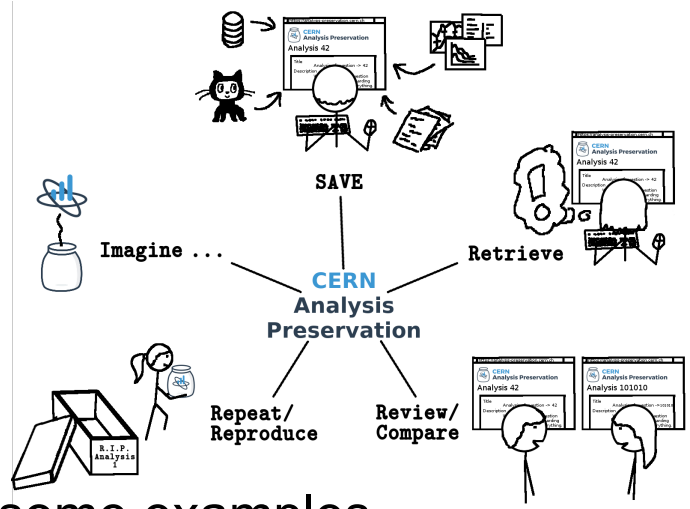
Henri Bachacou
Kyle Cranmer
Cristinel Diaconu
Roger Jones
Kevin Kroeninger
David Rousseau
David South (chair)
Physics Coordination (ex-officio)

Additional sources of information in the report

Frank Berghaus (ATLAS)
Sunje Dallmeier-Tiessen (CERN Scientific Information Service)
Robin Daser (CERN Scientific Information Service)
Lukas Heinrich (ATLAS)
Tibor Simko (CERN IT CDA-Digital Repositories)
Tim Smith (CERN IT Collaboration, Devices & Applications)

Analysis preservation: Use cases

- A long and varied list of use cases for Analysis Preservation appears in the report
  - A result of our discussions within the panel but also together the CERN Analysis Preservation (CAP) group. Their own list is here: https://github.com/cernanalysispreservation/analysispreservation.cern.ch/wiki/Use-cases



- The complete list is in the back-up, here are just some examples
  - **To address questions or concerns about published analyses, internal or otherwise**
  - **To assist in knowledge transfer if a person leaves the collaboration and has to hand over the know-how to other members**
  - **Make it easier for someone new to join a physics analysis and/or a CP group**
  - **To allow reinterpretation of an existing analysis for (e.g.) a new model of BSM physics**
  - **To allow an analysis to be repeated, which may be desirable due to improved precision of newly available theoretical predictions**
  - **To facilitate the combination of similar data from different experiments**
  - **Provide the possibility to re-run the actual code that was used for the analysis**
  - **A way to prepare an analysis for outreach activities such as the event challenge**

Strategy 1. ATLAS analysis (meta)data consolidation

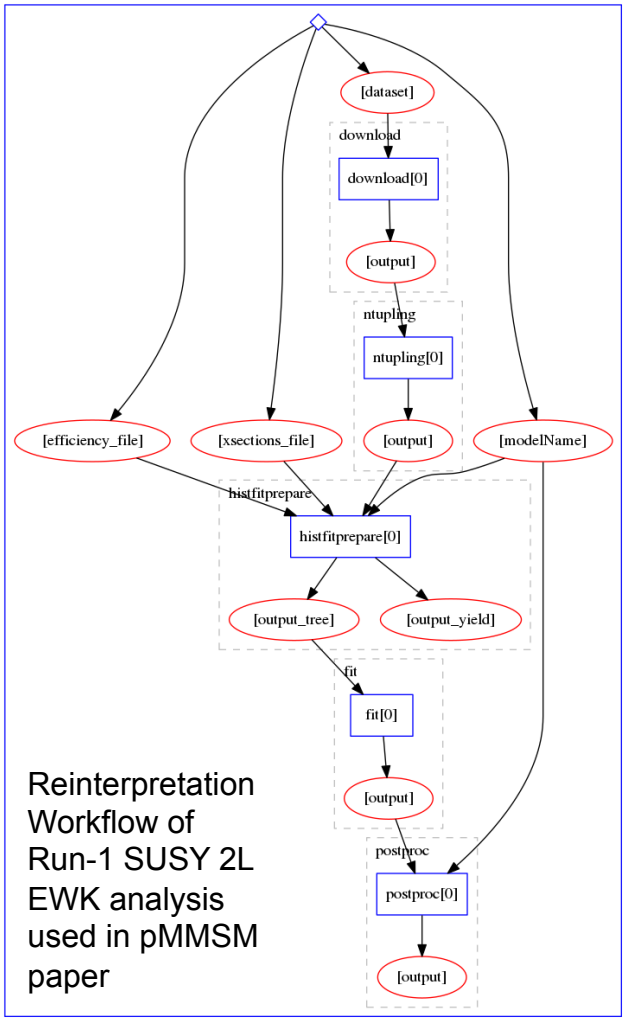
- The majority of the metadata about an analysis in ATLAS is to be found either on Glance or in AMI - and soon to be in AMI-Glance
  - Additional metadata sources include JIRA, twiki, svn/GIT, CDS, indico,..
  - And it's up to ATLAS to ensure all relevant information is in AMI-Glance

- Initial efforts with the CAP required manual filling of fields in a web interface
  - If this information can then also be exported from AMI-Glance as json/xml, this can be harvested by the CAP interface, which can fill this information into the correct fields
  - Again, this requires ATLAS to design its CAP interface together with the CAP team
  - E.g., datasets listed in AMI-Glance contain AMI tag, which is used to harvest specific info for CAP
  - Two way exchange, retrieval and upload, also possible with resources such as HEPDATA

- Preservation of associated analysis level data is also foreseen, with initial estimates of 1 TB per analysis, to be ingested from for example EOS
  - The CAP foresees search functionality and would also recognise duplicate information, allowing commonalities between analyses to be identified

Strategy 2. Capture of analysis stages and workflow

- On-going integration work with CAP and DASPOS to develop schemas to capture and describe analyses, such that they can be re-run/re-used by the collaboration at a later stage
- More details in presentation by Lukas Heinrich at CHEP16
  https://indico.cern.ch/event/505613/contributions/2227701/attachments/1350847/2047270/Oral-v3-559.pdf
- Two-step process:
  - **A. Capture individual analysis stages** (event selection, fitting, ..)
    1) Analysis-specific software
    Parameterised capture of e.g. command line programs etc.., such that they can be re-run. Custom (RootCore/Athena) packages, small scripts, etc..
    2) Required software environments
    Complete description of required underlying software requirements such as system libraries, analysis releases via e.g. CVMFS, etc… description of authentication requirements (Kerberos, X509, etc..)
  - **B. Capture workflow that connects stages**
    1) Order of analysis stages
    e.g. *first* event selection, *then* limit setting
    2) Data dependencies
    e.g. for a given step what data from upstream stages is needed as input
- Details of experiences with workflows/CAP in the back-up

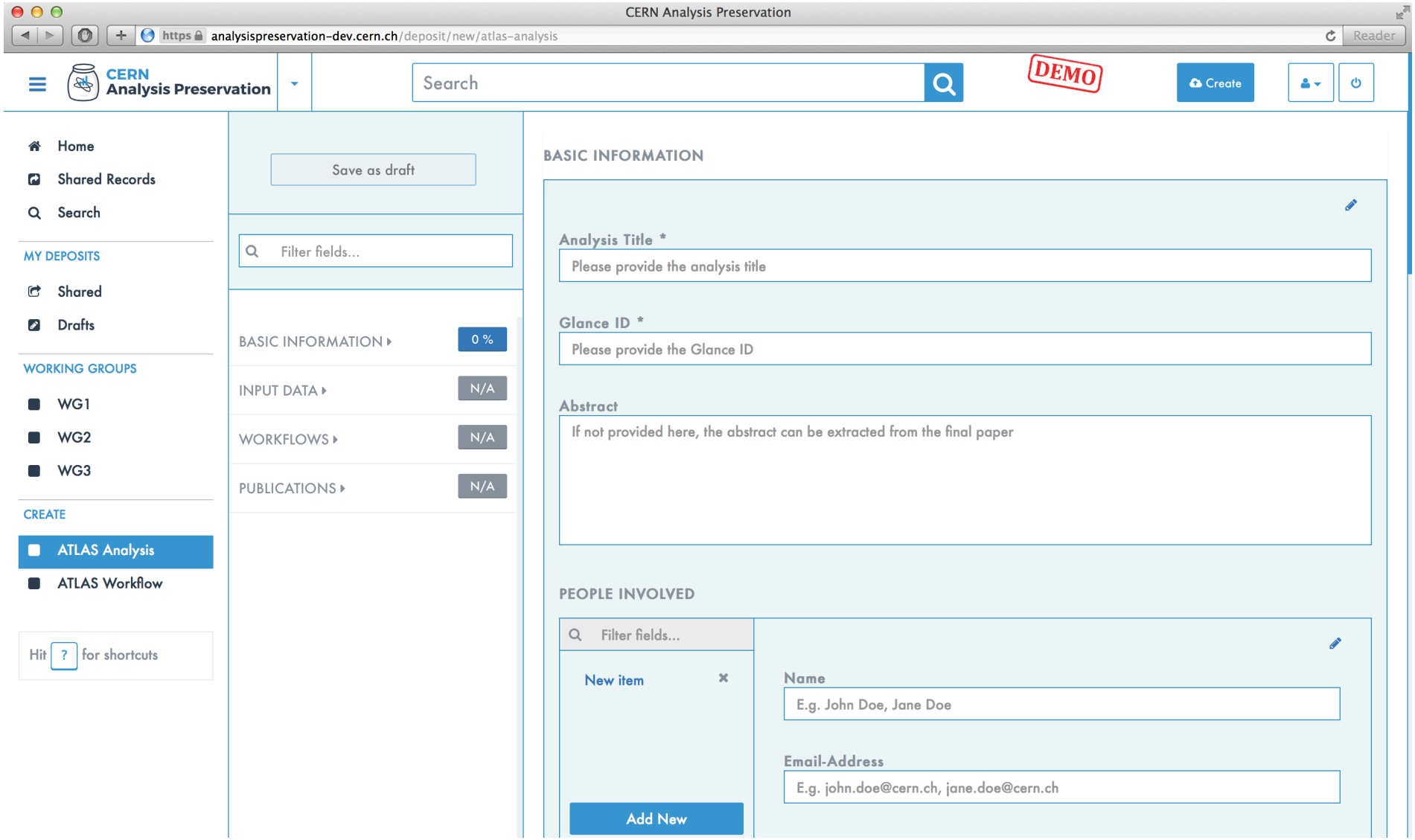


Reinterpretation Workflow of Run-1 SUSY 2L EWK analysis used in pMMSM paper

CERN Analysis Preservation

ATLAS Analysis Preservation in the CAP

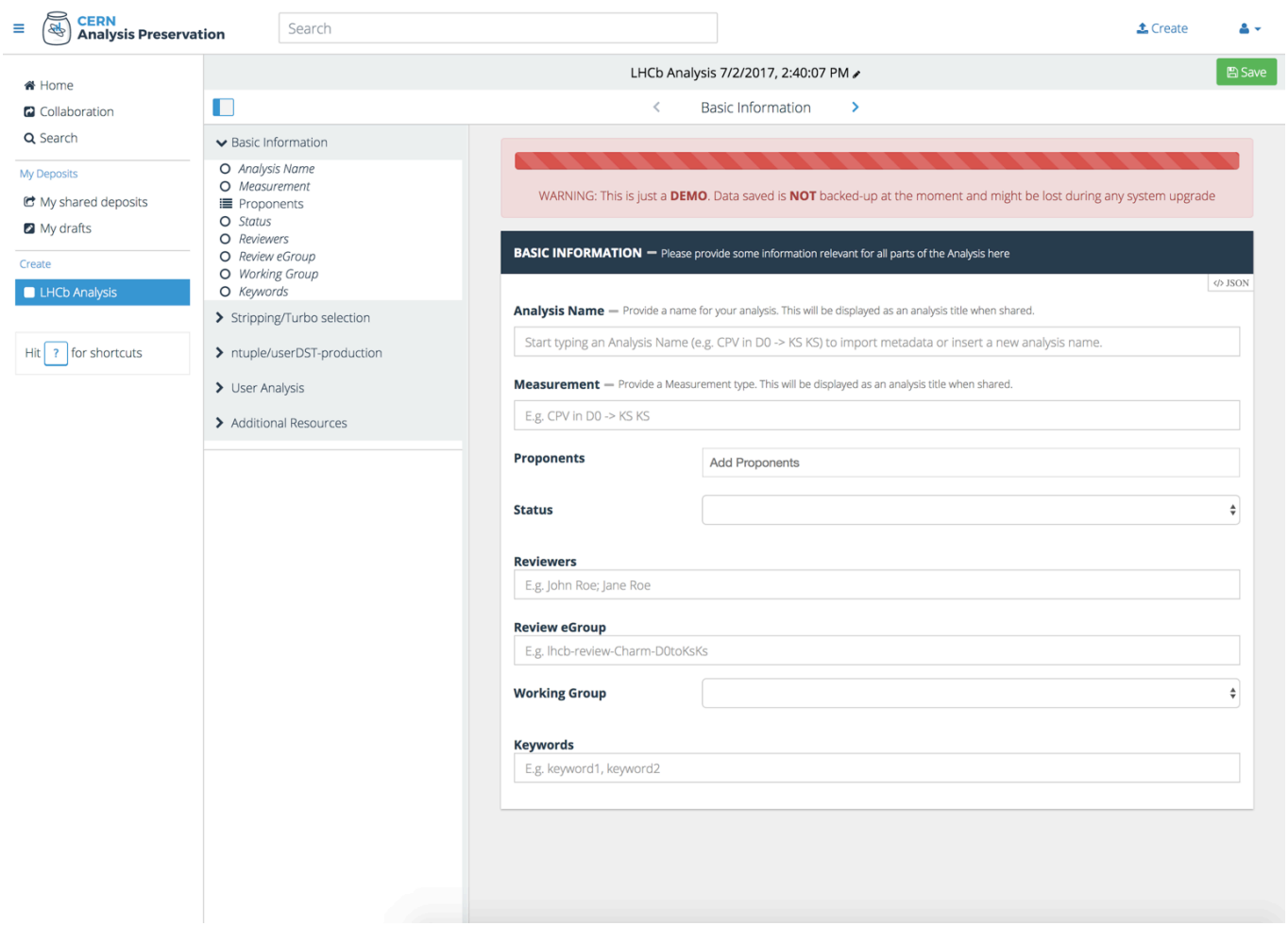
https://analysispreservation.cern.ch/deposit/new/atlas-analysis



LHCb interface in the CAP

How it works I

Submission via submission form or shell based ”client” using the API

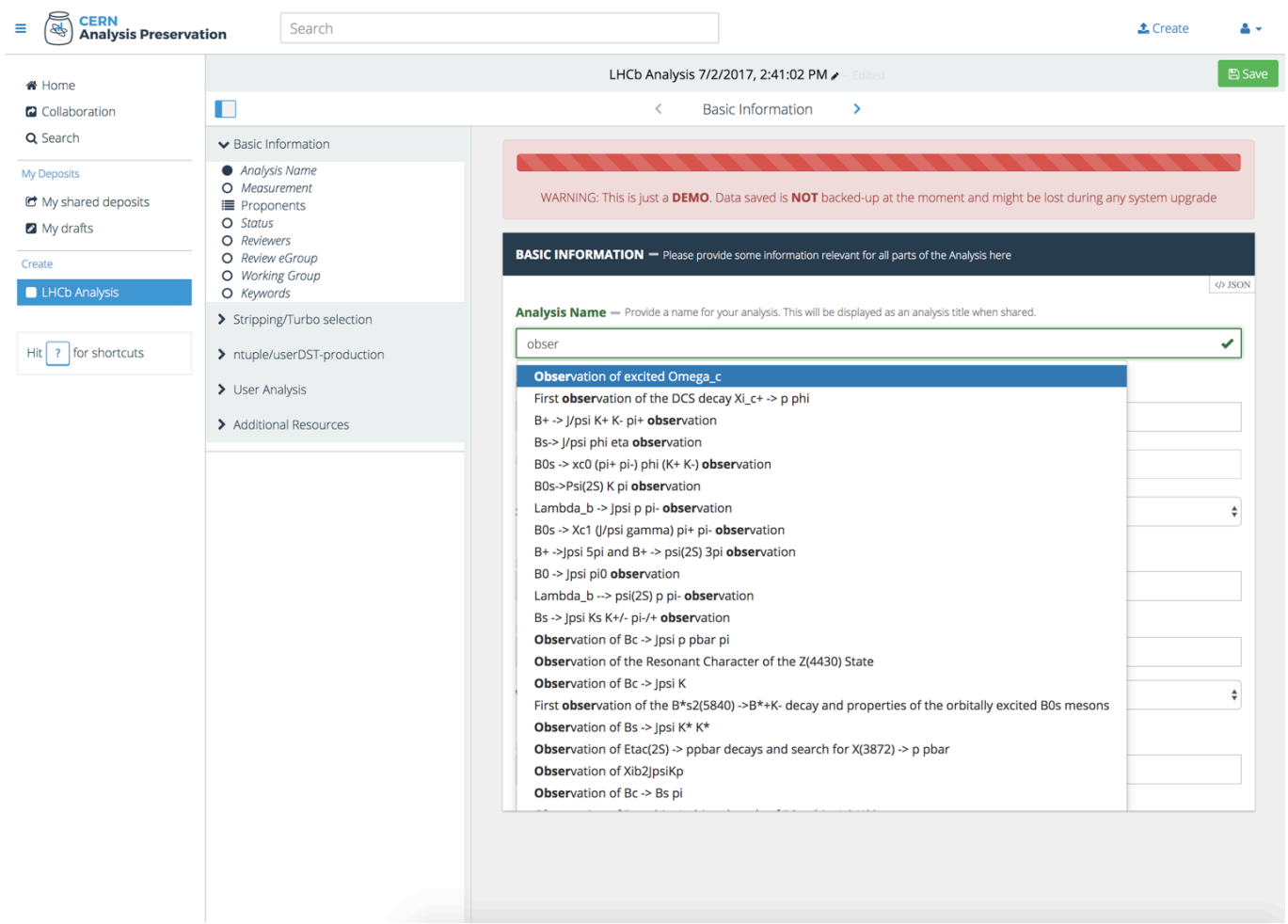


LHCb interface in the CAP

How it works II

Auto-complete functionality based on connections made to all relevant (existing) LHCb databases

Same possible for GLANCE, AMI and…



LHCb interface in the CAP

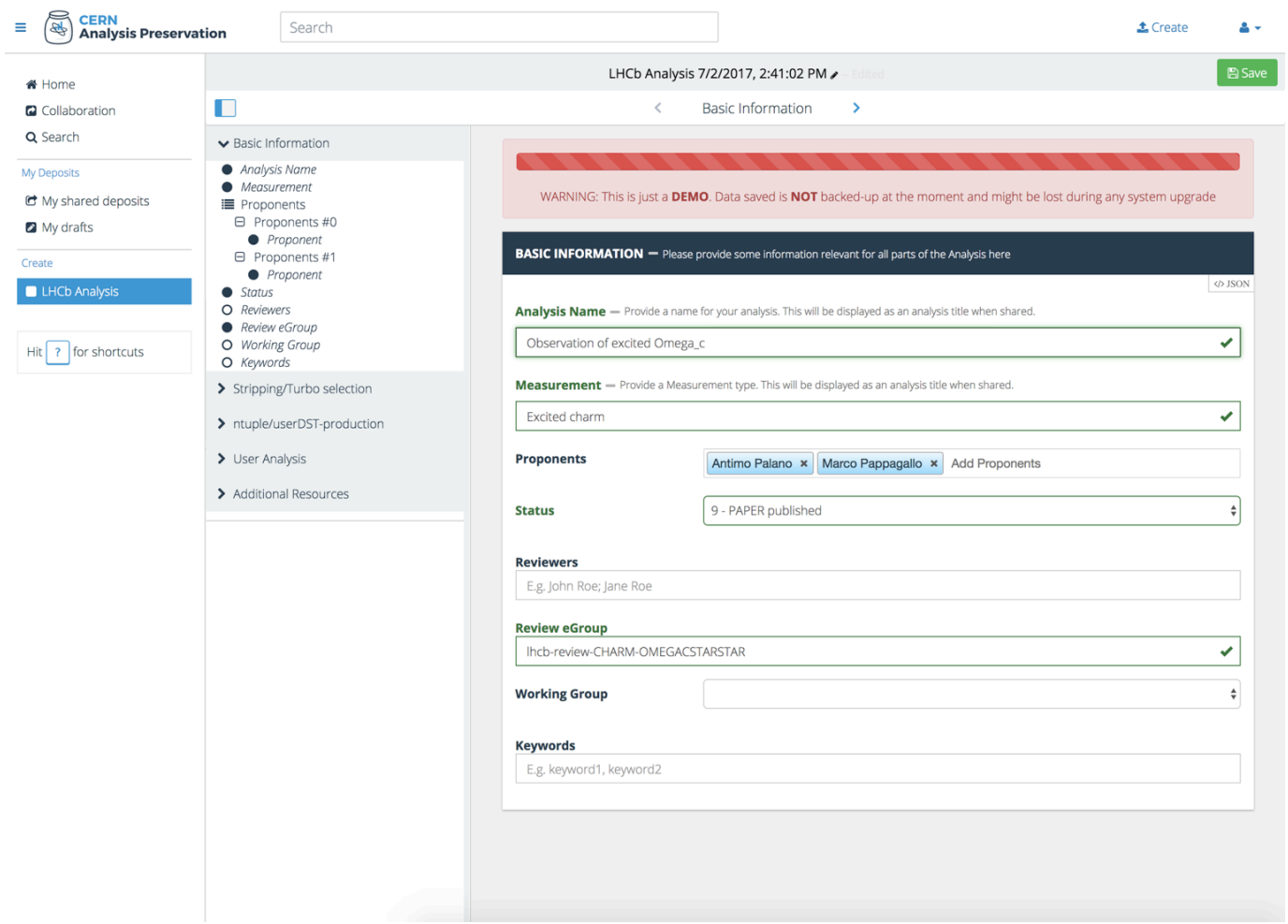
Physics metadata

Access to internal databases provides key information without much additional effort

This concerns:
Basic information e.g. authors, title and unique ID/name of an analysis

Further information, such as OS, analysis software and related internal discussions, presentations and publications

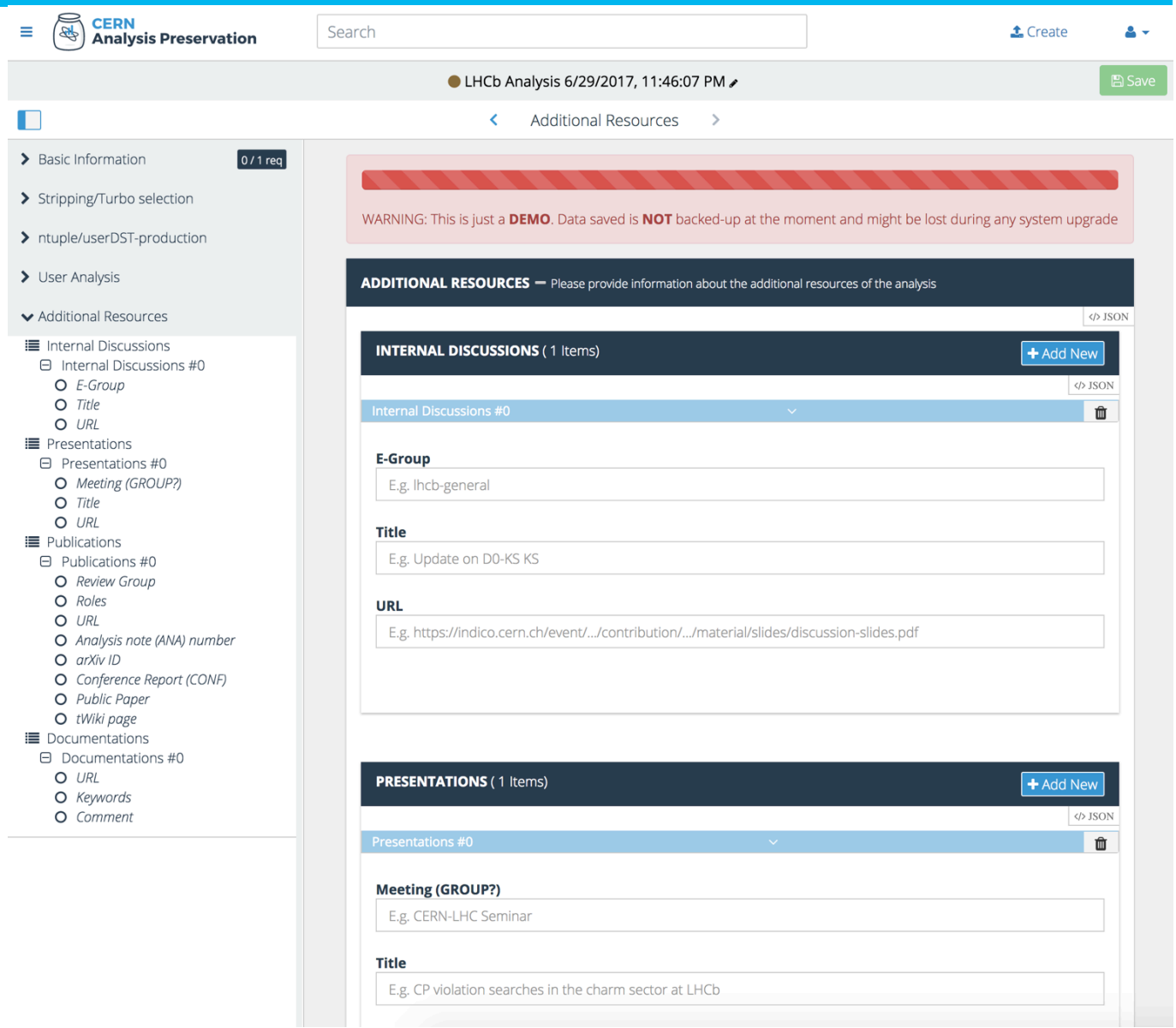
Detailed physics information (e.g. final state particles, cuts and vetos) for future reuse



LHCb interface in the CAP

Content from/ for you

Connecting and capturing internal notes, presentations, publications, documentation, etc
This will also include Twiki pages.

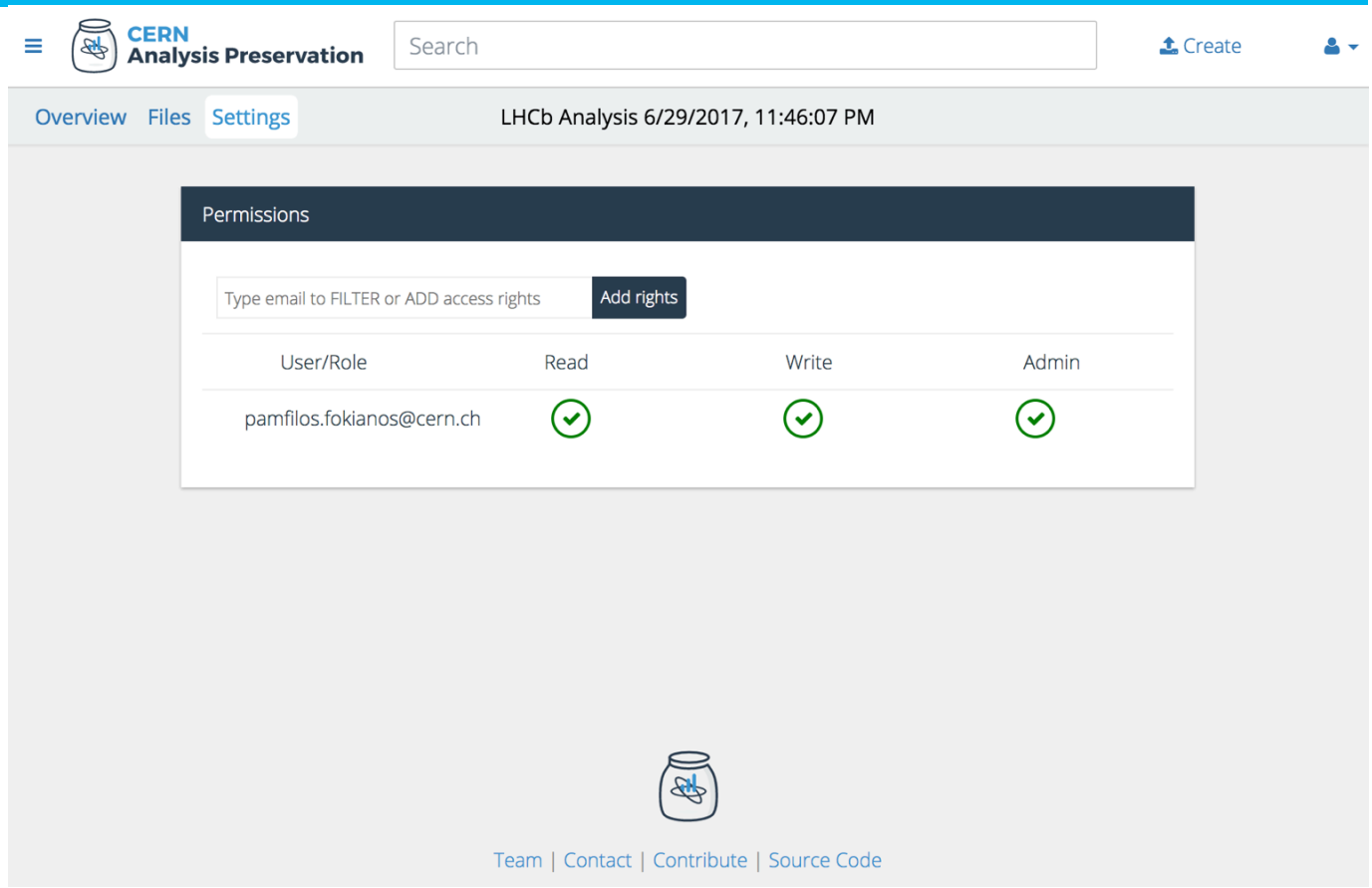


LHCb interface in the CAP

Permissions

Standard access restrictions apply (CERN e-groups)

For editing and contributing researchers can assign more refined permissions to their submissions



Extra material

Analysis preservation: Full list of use cases

- A complete description of how an analysis was performed, capturing all relevant information and metadata is desirable in order to:
  - **Address questions or concerns about published analyses, internal or otherwise**
  - **Identify who has worked on a particular data set, piece of software or Monte Carlo production**
  - **Assist in knowledge transfer if a person leaves the collaboration and has to hand over the know-how to other members**
  - **Make it easier for someone new to join a physics analysis and/or a CP group**
  - **Streamline moving to new versions of data and/or software**
  - **Provide the source for a comprehensive set of metadata concerning a presentation or publication which is to be submitted for internal review**

Analysis preservation: Full list of use cases

- Preserving the knowledge and details of how an analysis was performed, allows *Reproducibility* and/or *Reusability* in many ways
  - **Reinterpretation of an existing analysis for (e.g.) a new model of BSM physics**
  - **An analysis to be repeated, which may be desirable due to improved precision of newly available theoretical predictions**
  - **The extension of an analysis with new data**
  - **Easier comparisons to be performed of work done by two groups working independently within the collaboration**
  - **A conflict between results of two different collaborations on the same subject to be investigated and/or resolved**
  - **To facilitate the combination of similar data from different experiments**
  - **To allow model interpretations to enter the publication process later**
- Depending on the implementation, Analysis Preservation may also provide:
  - **The possibility to re-run the actual code that was used for the analysis**
  - **A tool to reproduce analyses semi-automatically, so that WGs can plug their analyses into it**
  - **A way to prepare an analysis for outreach activities such as the event challenge**
  - **An easier method for preparing the data for open access**

Strategy 1. ATLAS analysis (meta)data survey

	Structured, Public	Structured, Internal	Unstructured, Internal
Documentation	- Publications (arXiv) - Proceedings (INSPIRE) - Conference talks (indico, www) - Master theses, PhDs, and Habilitations (www)	- Personnel, workflow, structures (Glance) - Internal notes (CDS) - Approval talks (various twikis)	- Paper comments, early drafts (CDS) - Editorial Board meetings (indico) - Group/subgroup meetings (indico) - eGroup discussions - JIRA tickets - twiki pages - Private email threads
Software	- Monte Carlo generators	- Releases and configuration of central production, as well as generation, simulation, reconstruction and derivations (AMI) - Tagged version(s) of group/user software for final ntuples and/or derivations (svn/GIT) - Tagged version(s) of code used to perform all components of the final analysis (svn/GIT)	- Earlier software versions (svn/GIT) - JIRA tickets - twiki pages - Private software
Data	- HEPDATA tables	- Datasets from central production: xAOD, derivations - Group ntuples and their derivations - Final histograms used for plots in the publication - Systematics data - Workspaces	- Earlier versions of the relevant datasets - Failed attempts

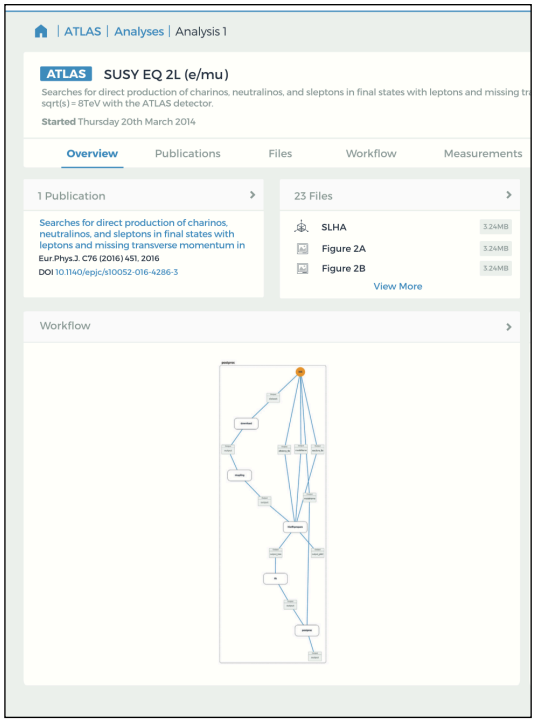
Strategy 2. Capture of analysis stages and workflow

> Experience

- Try to minimize burden on analysis teams. Information can be provided in human-readable text files, closely resembles existing work-patterns
- Very good experience with usage of Linux Containers (Docker etc) in the context of analysis software. **Allows us to heterogeneous analysis setups to be captured in a portable way**. Expect further simplification through rapid adoption of GitLab service and cmake migration
- Working examples for ATLAS full chain (run1/run2), SUSY analyses (run 1 electroweak, run 2 multi-b-jet)
- Workflows simplified during run-2 versus run-1 thanks to more centralised computing (derivation framework): easier capture
- Captured information stored in human/machine-readable JSON format, natively archived by CAP: indexable, searchable, composable (multiple analyses sharing common analysis stages)
- Workflow system and prototype deployment of re-execution cluster on CERN infrastructure used to provide results for parts of ATLAS-internal reinterpretation campaigns covering O(1000) new model points. (see: https://arxiv.org/abs/1508.06608, https://arxiv.org/abs/1608.00872)

```
process:
  process_type: 'interpolated-script-cmd'
  script: |
    #!/bin/bash
    source ~/.bashrc
    setupATLAS
    source ./rcSetup.sh
    /recast_auth/getmyproxy.sh
    lsetup fax dq2
    python MultibjetsAnalysis/scripts/Run.py --dataSource 1 --doSyst 1 --doNTUPSyst 1 --doNTUP 0
    mv {submitdir}/data-output_histfitter/*.root {outputprefix}.{did}.root
publisher:
  publisher_type: 'fromglob-pub'
  globexpression: '*.root'
  outputkey: histfitterfile
environment:
  environment_type: 'docker-encapsulated'
  image: lukasheinrich/multibsel_cvmfs
  resources:
    - CVMFS
    - GRIDProxy
```

Information needed to capture Run-2 SUSY analysis event Selection stage (DxAOD → HistFitter)



Integration of Workflows into CERN Analysis Preservation

CERN Analysis Preservation