PAMFILOS FOKIANOS

CERN ANALYSIS PRESERVATION

IMPROVE PRESERVATION AND RE-USE OF RESEARCH RESULTS AT CERN

CERN ANALYSIS PRESERVATION

Tool for physicists to capture, preserve and make their analysis reusable for the future

Development: CERN SIS-OS

Collaboration/Consulting: Tibor Simko, CERN IT, LHC experiments

OPEN SCIENCE

The accessibility, preservation and reusability of research outputs is an important challenge of our times. Many funders, political bodies and research organisations have stressed the importance of **Open Science**, and repeatedly argued for the value of **Open Data**.

Enabling openness and **reproducibility** through preserving the **entire research data analysis workflow** ensures that the value of physicists' work is retained many years in the future.

MAIN GOALS OF CERN ANALYSIS PRESERVATION

- Enable the High-Energy Physics community to capture, preserve and share their research objects (metadata, files, tables, plots, wikis, etc...)
- Offer a stable, flexible and collaborative tool for physicists to capture and share their work
- Accommodate the needs of each collaboration/team, depending on the specifications and requirements each of them have.
- Support open practices but ensure that users are always in control of when and if their work is shared or published.

Capture

Collect and preserve elements needed to understand and rerun your analysis

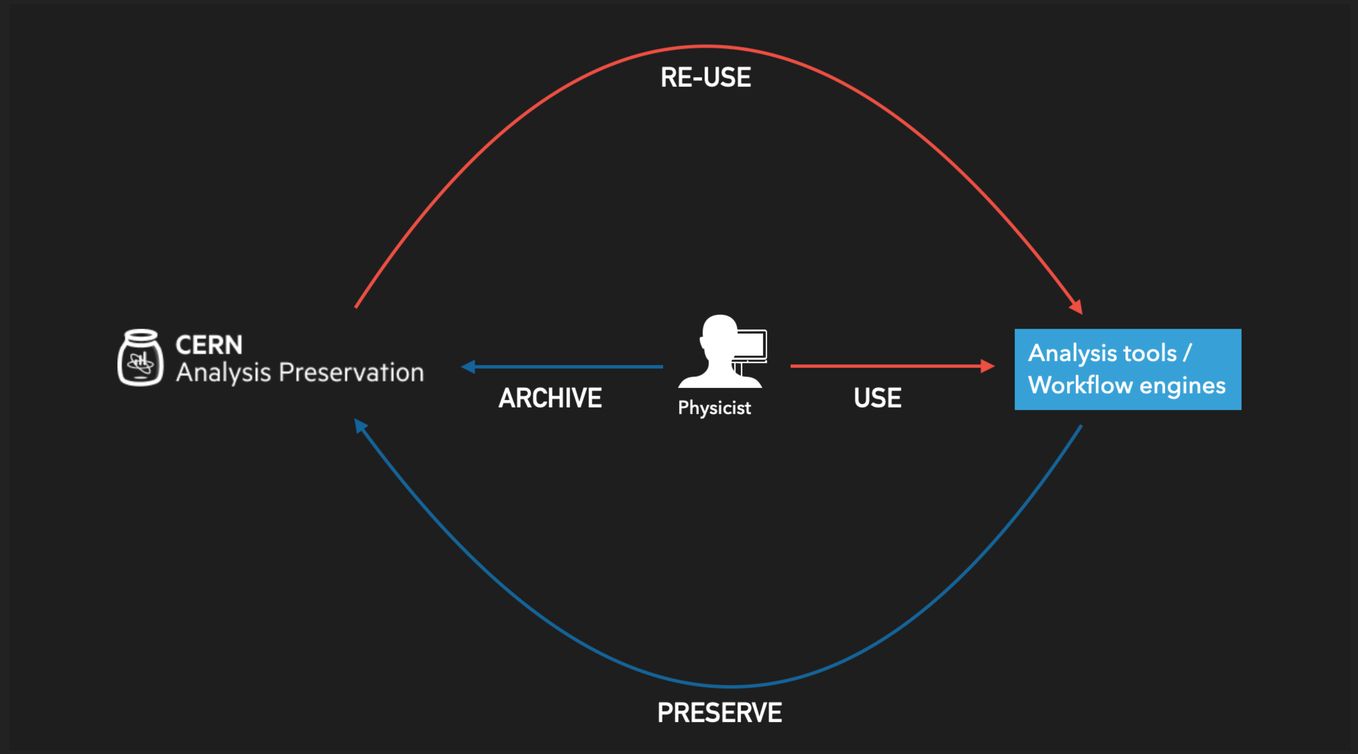
Collaborate

Share your analysis and components with other users, your collaboration or group

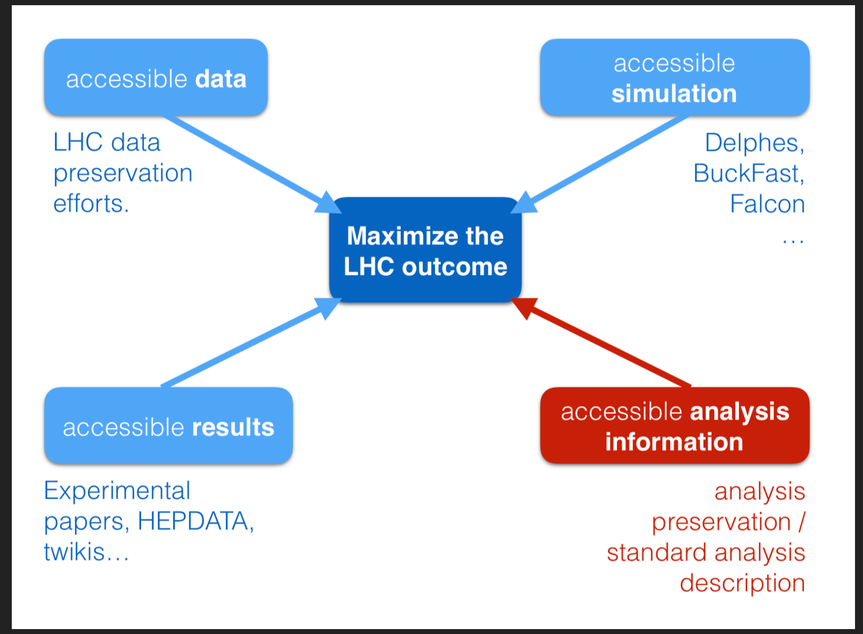
Reuse

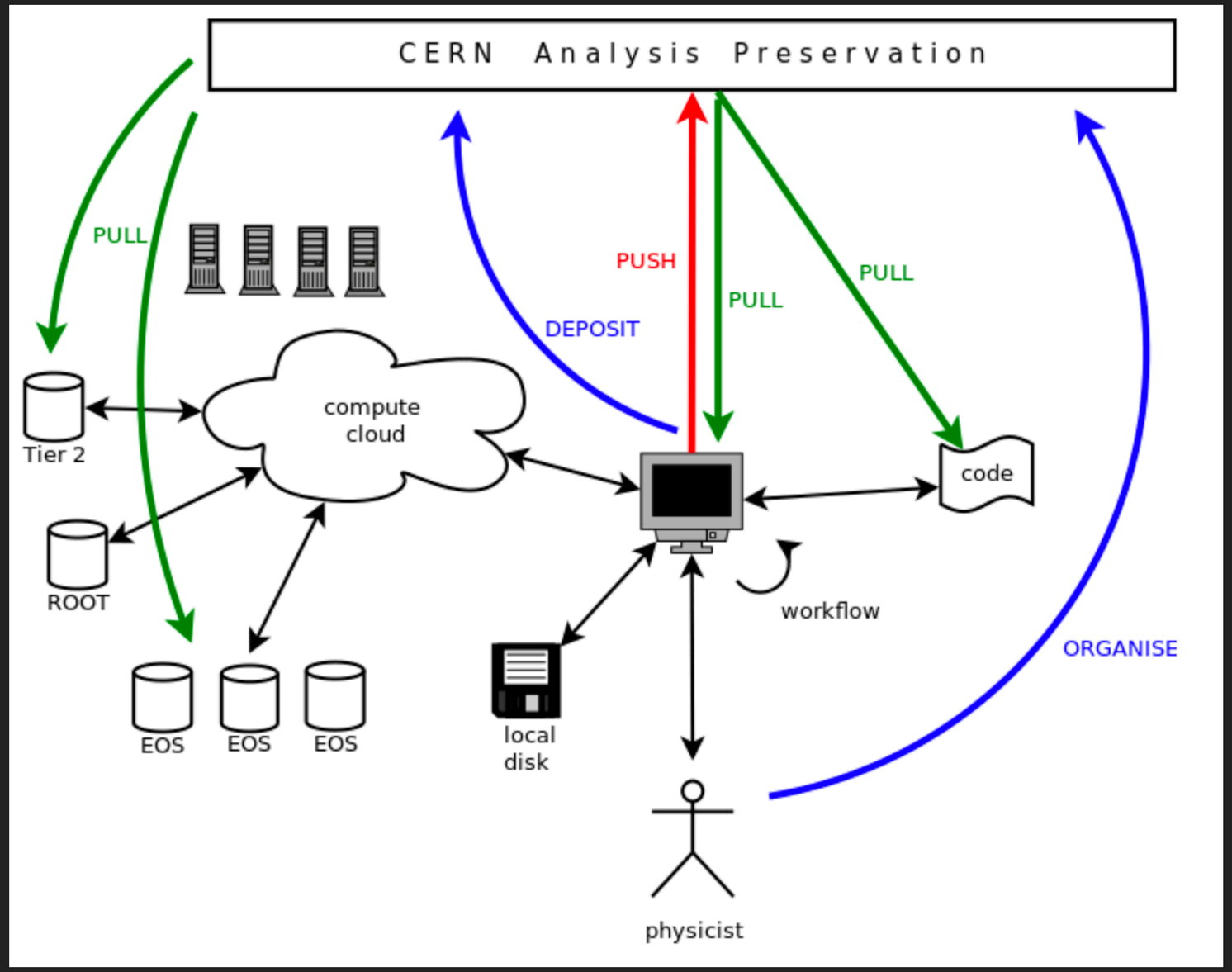
Easily reuse analysis components and help run containerised workflows

PRESERVATION ⇄ RE-USABILITY



MAXIMIZE LHC OUTCOME

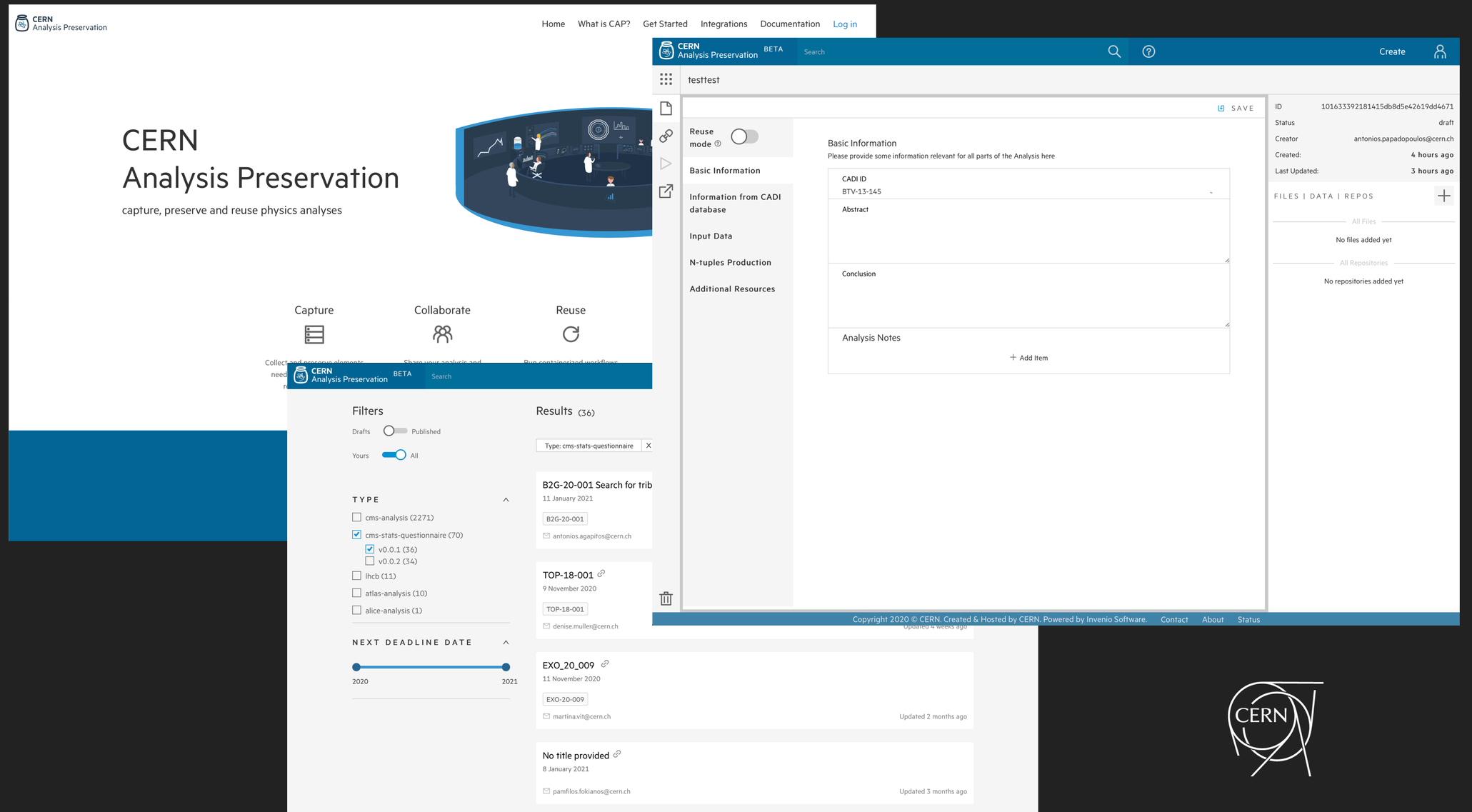



CERN Analysis Preservation
PULL
PUSH
PULL
PULL
DEPOSIT
compute cloud
Tier 2
ROOT
code
workflow
ORGANISE
EOS
EOS
EOS
local disk
physicist

WHAT WE PROVIDE

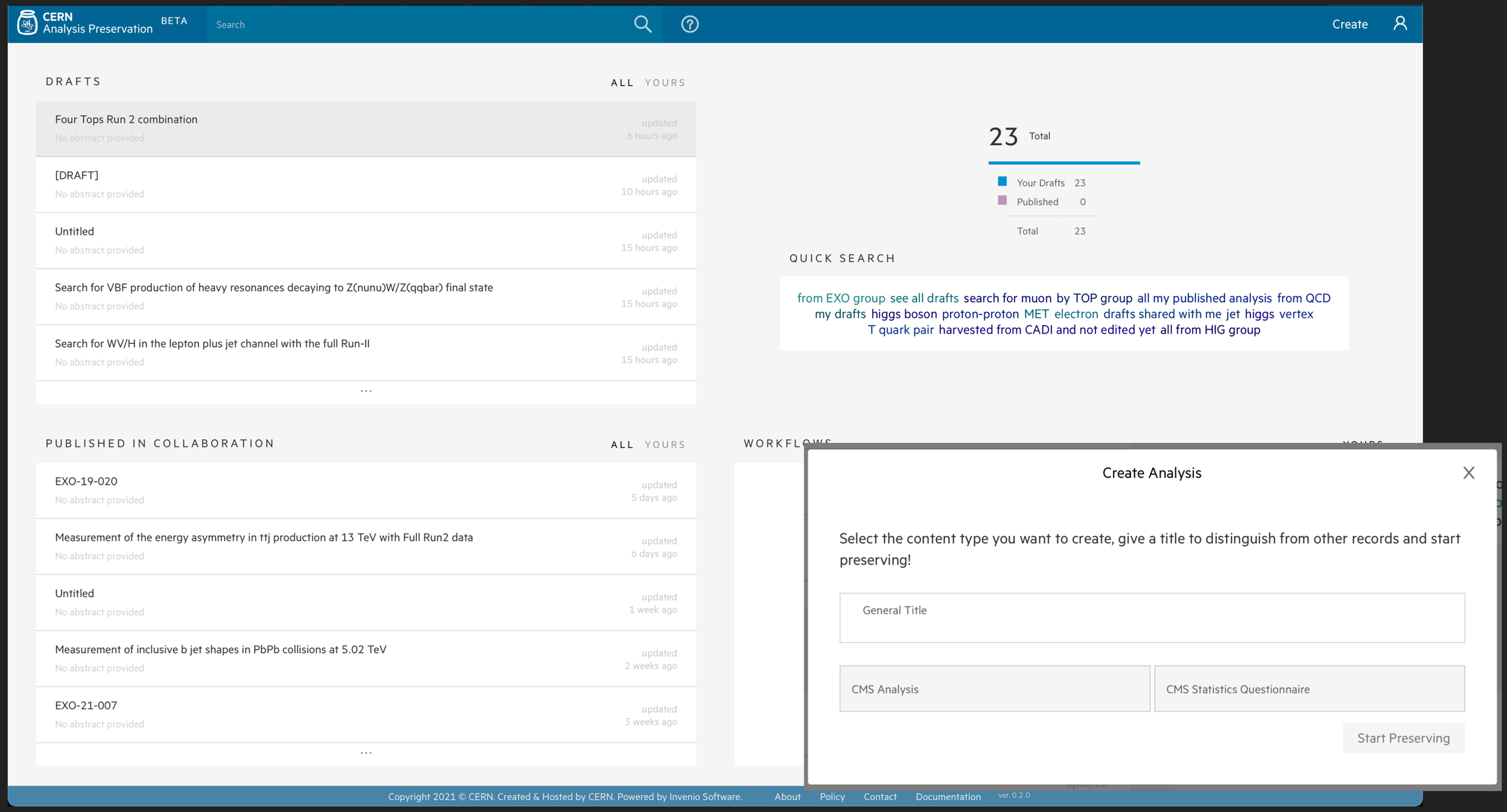
- Advanced and clean **forms**
- **Sync** and **index** with various **experiment Databases/APIs/sources** for searching and integration
- Easy, automated and efficient **file uploading**
- **Flexible data models** (JSON-Schemas), FAIR practices,
- **Versioning** of metadata & files, using the publishing draft/record model
- Integration with **related scientific services and persistent identifiers** (i.e. Github, Gitlab, Zenodo, ORCID, ROR, etc)
- Ongoing integration with CERN services that support remote execution and reuse (e.g. REANA) of **computational workflows**
- **Review, notifications** features
- **Powerful REST API** (records, files, users, search)
- Piloted in collaboration with all the major LHC experiments at CERN

CERN ANALYSIS PRESERVATION

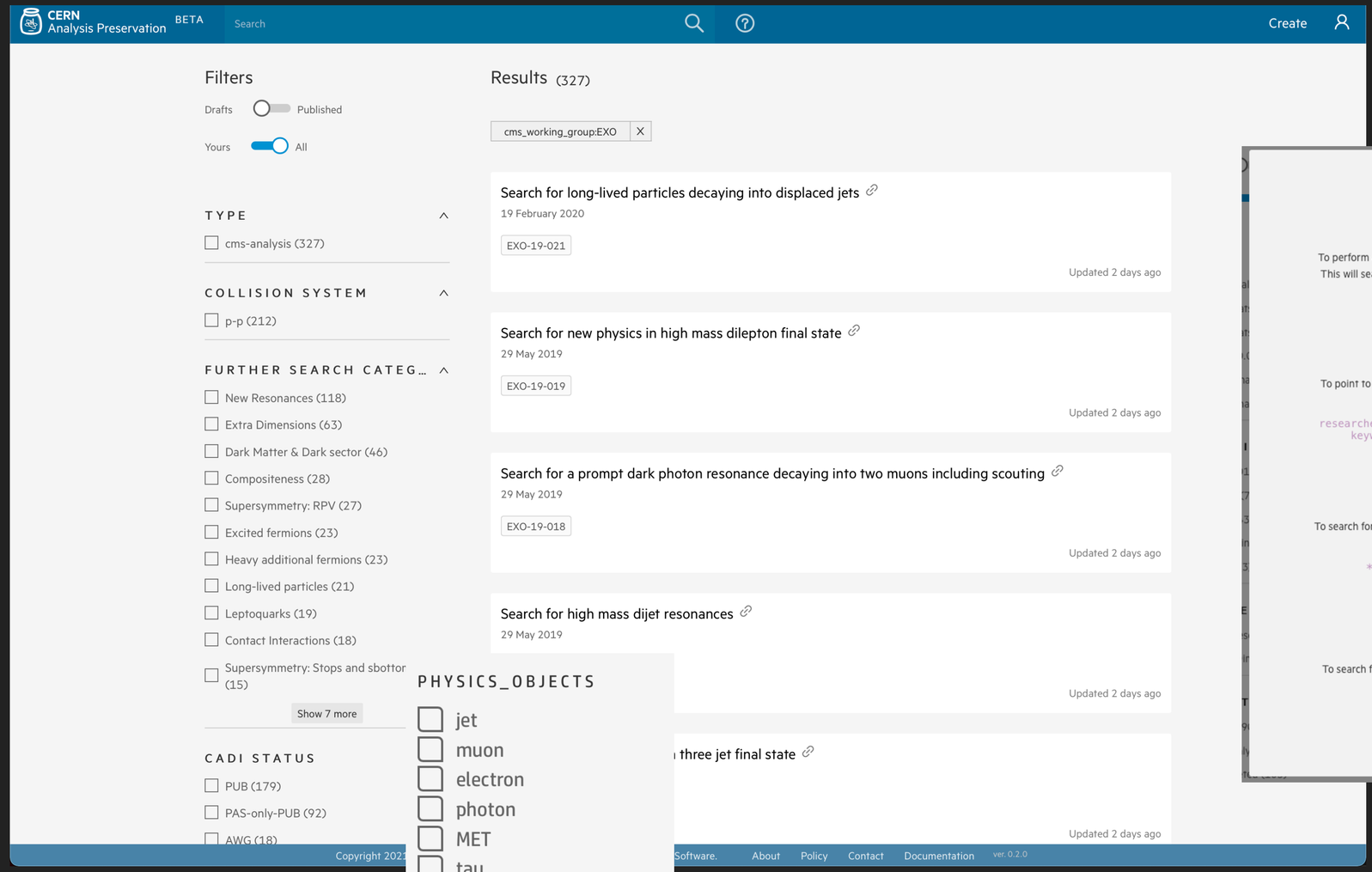


DASHBOARD

General overview of drafts/published, quick search, create new content



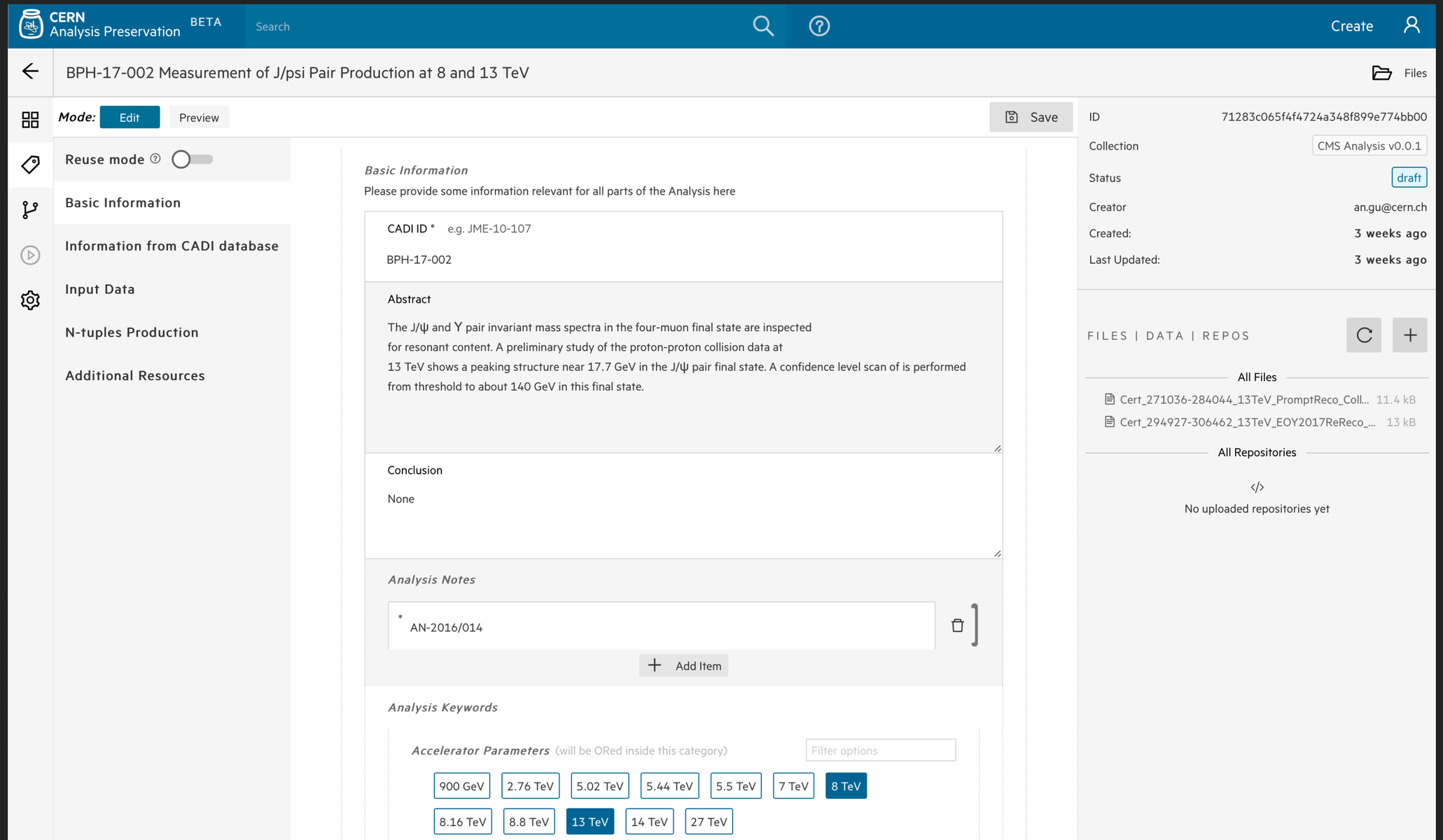
ADVANCED SEARCH



CERN

CONTENT/DRAFT EDITOR

Metadata editor



CONTENT/DRAFT EDITOR

Automated fetching from experiment DBs/APIs/other sources

CADI ID * e.g. JME-10-107

EXO-17-023

CMS CADI

Information from CADI database

Automatically taken from CADI, based on CADI ID

Name
Search for Black Holes and Sphalerons

Description
Search for Black Holes and Sphalerons (full 2016 data)

Contact Person
martin_ka_hei@yahoo.com.hk

Twiki
https://twiki.cern.ch/twiki/bin/view/CMS/EXO17023

Created
2017-10-03

Paper
http://cms.cern.ch:80/iCMS/analysisadmin/get?analysis=EXO-17-023-paper-v27.pdf

PAS
http://cms.cern.ch:80/iCMS/analysisadmin/get?analysis=EXO-17-023-pas-v1.pdf

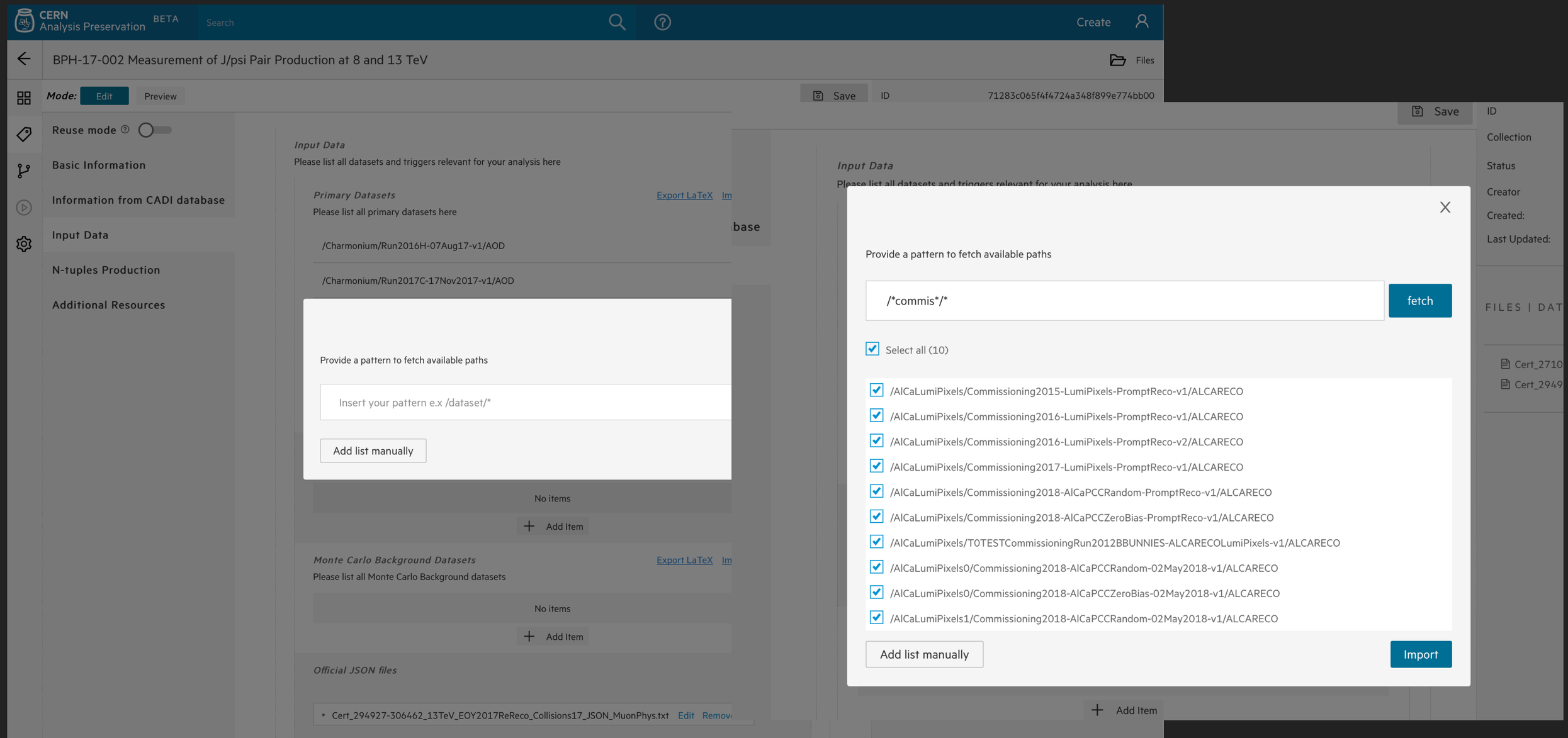
Publication Status
this field value will be automatically loaded when a valid CADI ID will be provided

Status
PUB

e.g. CADI integration/sync

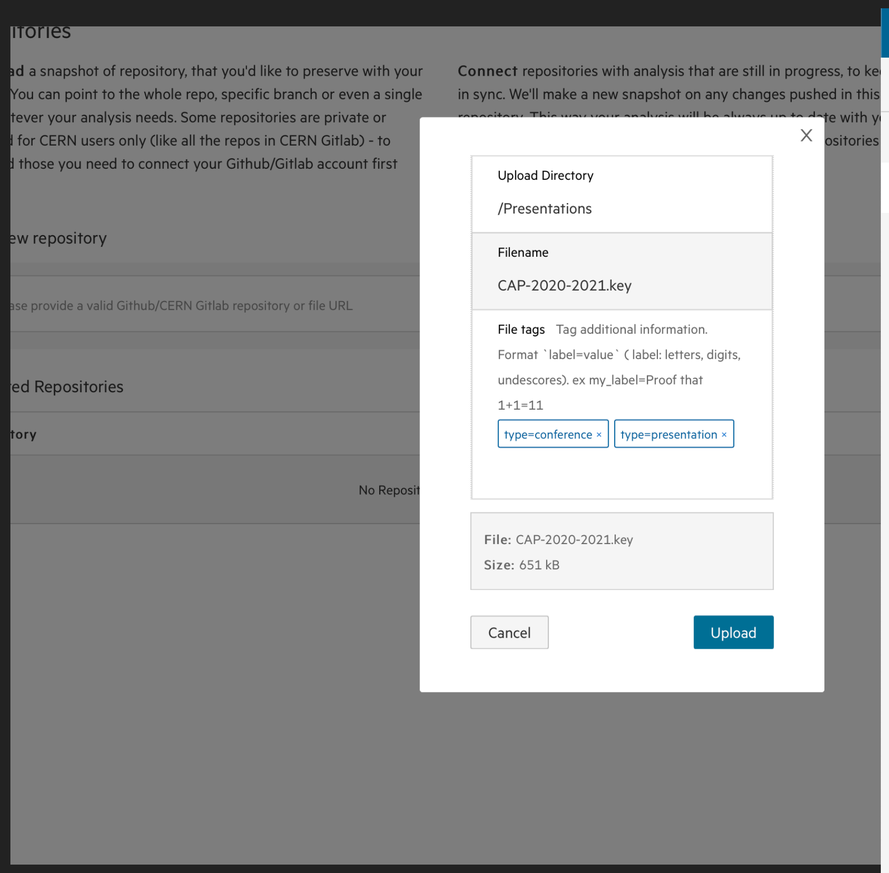
CONTENT/DRAFT EDITOR

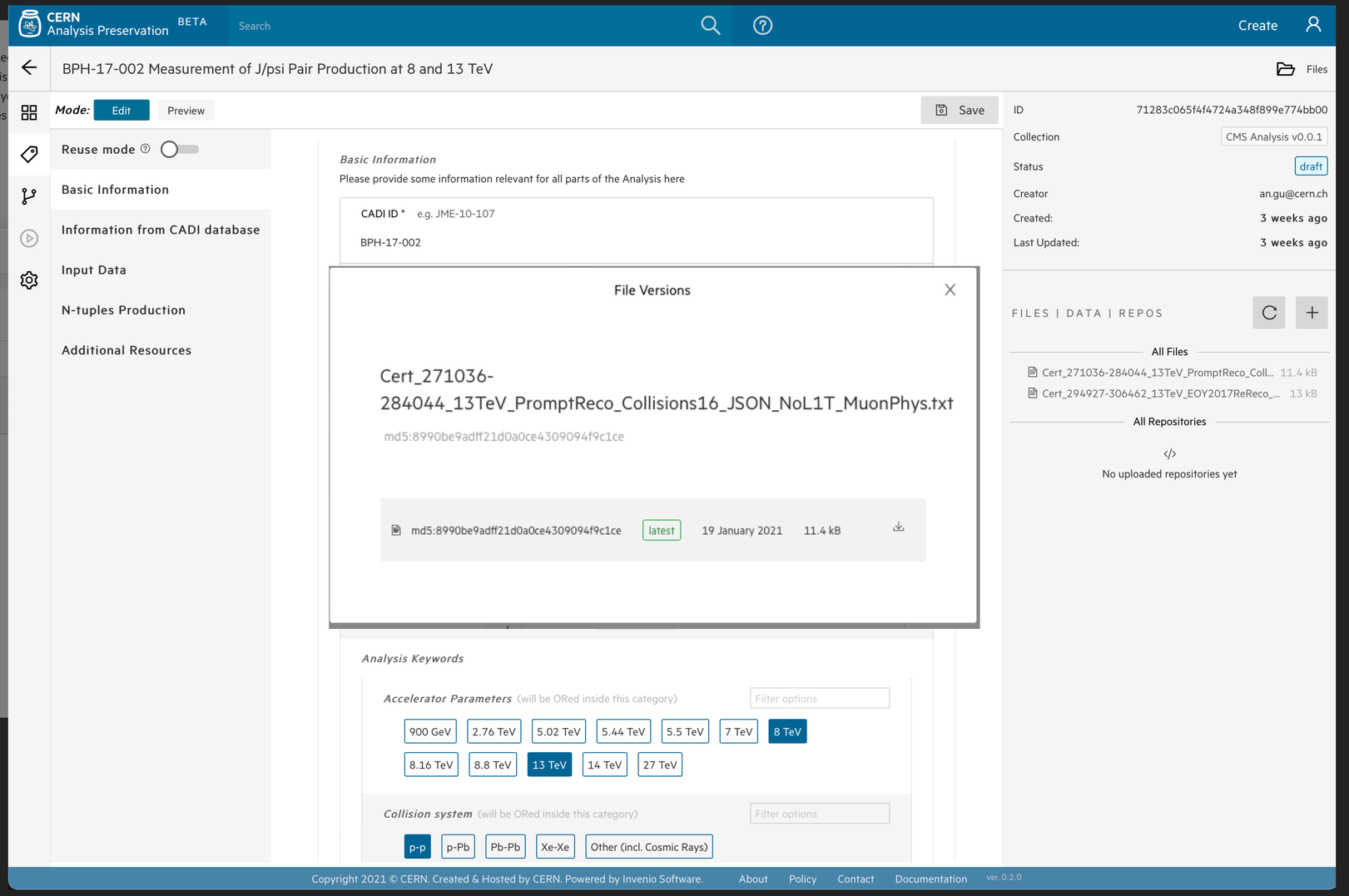
Suggestions & auto-completes



CONTENT/DRAFT EDITOR

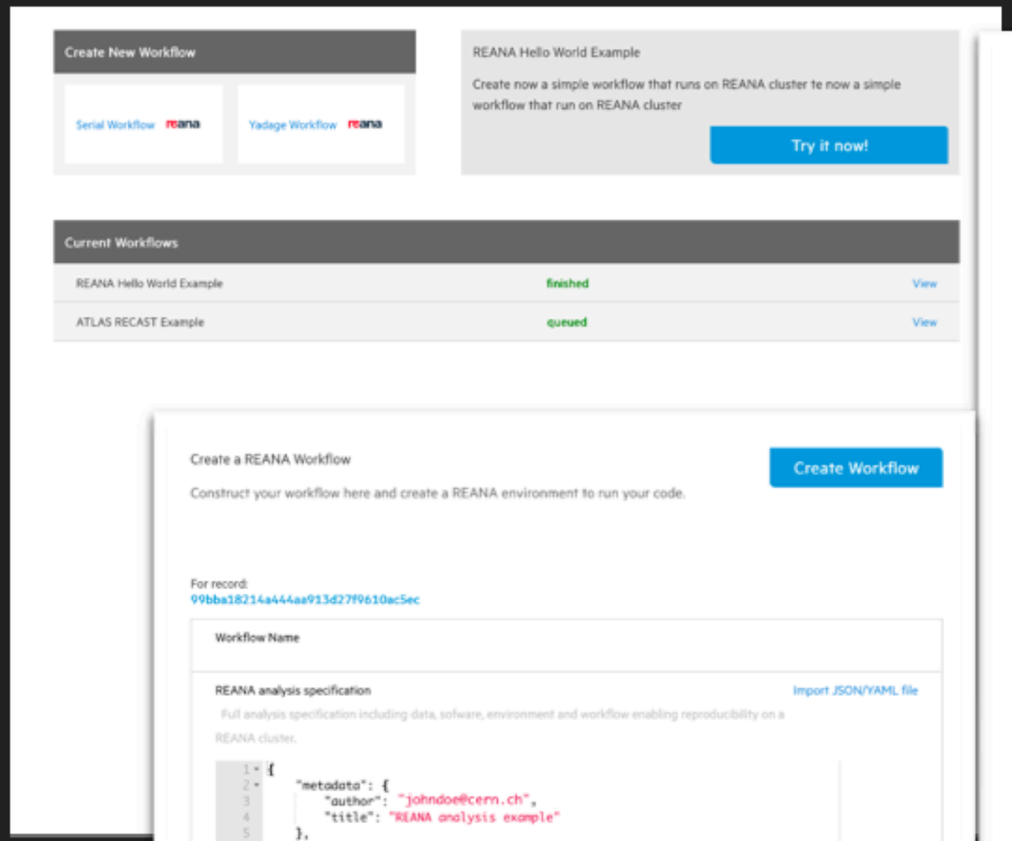
File upload - tagging, versioning

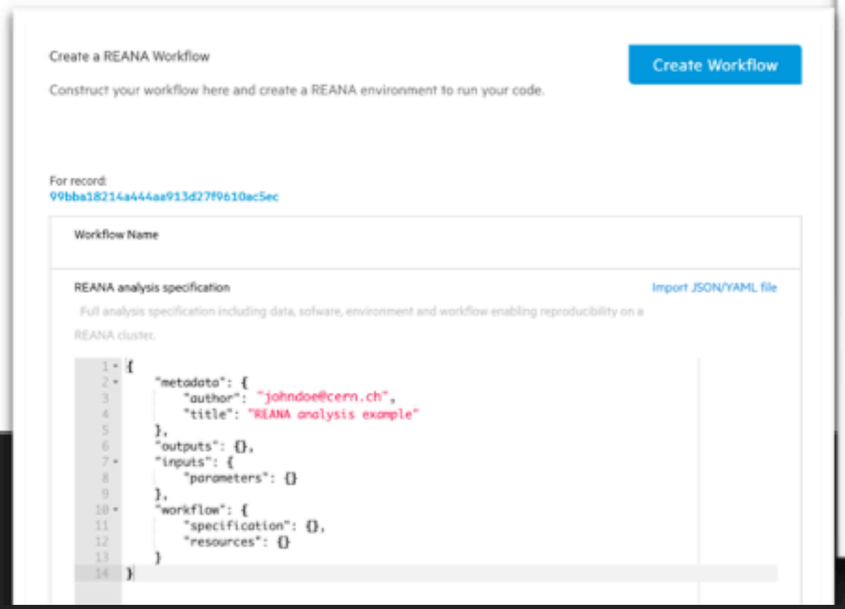


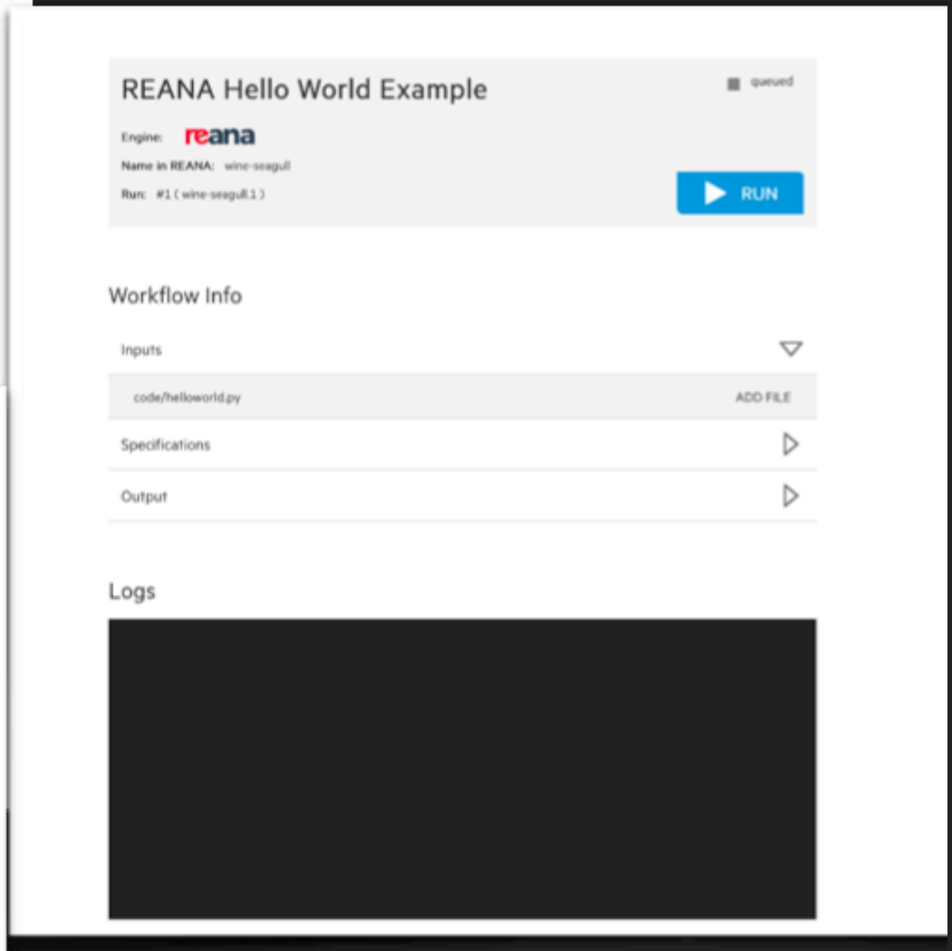


WORKFLOWS – REANA INTEGRATION

Use data, files and info from CAP to create a workflow in REANA - connect it with user analysis, check status, rerun with different parameters. Prototype / Not released yet







PUBLISHED CONTENT

Public to everyone in the collaboration

CERN Analysis Preservation BETA Search Create

Observation of BR(B+ --> psi(2S) phi K+)

Edit

Published | CAP.CMS.DYTU.HGFN | CMS Analysis v0.0.1

Reuse mode

Basic Information

Information from CADI database

Input Data

N-tuples Production

Additional Resources

Basic Information

CADI ID

Abstract

Conclusion

Analysis Notes

AN-2013/384

Analysis Keywords

Accelerator Parameters

Collision system

CERN Analysis Preservation BETA Search Create

Observation of BR(B+ --> psi(2S) phi K+)

Edit

Published | CAP.CMS.DYTU.HGFN | CMS Analysis v0.0.1

Contact Person	Kai Yi (IOWA)
Twiki	https://twiki.cern.ch/twiki/bin/viewauth/CMS/BPH13009
Created	26/11/2013
Paper	http://cms.cern.ch:80/iCMS/analysisadmin/get?analysis=BPH-13-009-paper-v34.pdf
PAS	http://cms.cern.ch:80/iCMS/analysisadmin/get?analysis=BPH-13-009-pas-v7.pdf
Publication Status	Free
Status	PUB

Input Data

Primary Datasets Export LaTeX

Please list all primary datasets here

/DoubleMuParked/Run2012A-22Jan2013-v1/AOD

/MuOniaParked/Run2012B-22Jan2013-v1/AOD

/MuOniaParked/Run2012C-22Jan2013-v1/AOD

/MuOniaParked/Run2012D-22Jan2013-v1/AOD

Monte Carlo Signal Datasets Export LaTeX

Please list all Monte Carlo Signal datasets

/BuToPsiPhiK BFilter TuneZ2star 8TeV – pythia6 – evtgen/ 219 Summer12 DR53X – PU RD2 START53 V19F – v2/AODSIM

/BuToPsiK BFilter TuneZ2star SVS 8TeV – pythia6 – evtgen/Summer12 221 DR53X – PU RD2 START53 V19F – v2/AODSIM

Files | Data | Source Code

All Files

No uploaded files yet

All Repositories

No uploaded repositories yet

PUBLISHED CONTENT

Features focused on usability - e.g LaTeX exporter

CERN Analysis Preservation BETA | Search | Create

Observation of BR(B+ --> psi(2S) phi K+)

Edit

Published | CAP.CMS.DYTU.HGFN | CMS Analysis v0.0.1

Contact Person	Kai Yi (IOWA)
Twiki	https://twiki.cern.ch/twiki/bin/viewauth/CMS/BPH13009
Created	26/11/2013
Paper	http://cms.cern.ch:80/iCMS/analysisadmin/get?analysis=BPH-13-009-paper-v34.pdf
PAS	http://cms.cern.ch:80/iCMS/analysisadmin/get?analysis=BPH-13-009-pas-v7.pdf
Publication Status	Free
Status	PUB

Input Data

Primary Datasets — Export LaTeX

Please list all primary datasets here

- /DoubleMuParked/Run2012A-22Jan2013-v1/AOD
- /MuOniaParked/Run2012B-22Jan2013-v1/AOD
- /MuOniaParked/Run2012C-22Jan2013-v1/AOD
- /MuOniaParked/Run2012D-22Jan2013-v1/AOD

Monte Carlo Signal Datasets — Export LaTeX

Please list all Monte Carlo Signal datasets

- /BuToPsiPhiK BFilter TuneZ2star 8TeV – pythia6 – evtgen/ 219 Summer12 DR53X – PU RD2 START53 V19F – v2/AODSIM

Files | Data | Source Code

All Files

No uploaded files yet

13-009-paper-v34.pdf

LaTeX Code

```
\usepackage{tabularx}

\begin{tabularx}{ 10cm }{|X|}
\hline

    \textbf{Primary Datasets} \\ \hline
    /DoubleMuParked/Run2012A-22Jan2013-v1/AOD \\ \hline
    /MuOniaParked/Run2012B-22Jan2013-v1/AOD \\ \hline
    /MuOniaParked/Run2012C-22Jan2013-v1/AOD \\ \hline
    /MuOniaParked/Run2012D-22Jan2013-v1/AOD \\ \hline

\hline
```

Copy to Clipboard

CAP CLIENT

Seamless and easy contributions for users work

From services, scripted or with CI

Push / pull metadata, code, files

```
$ pip install cap-client
```

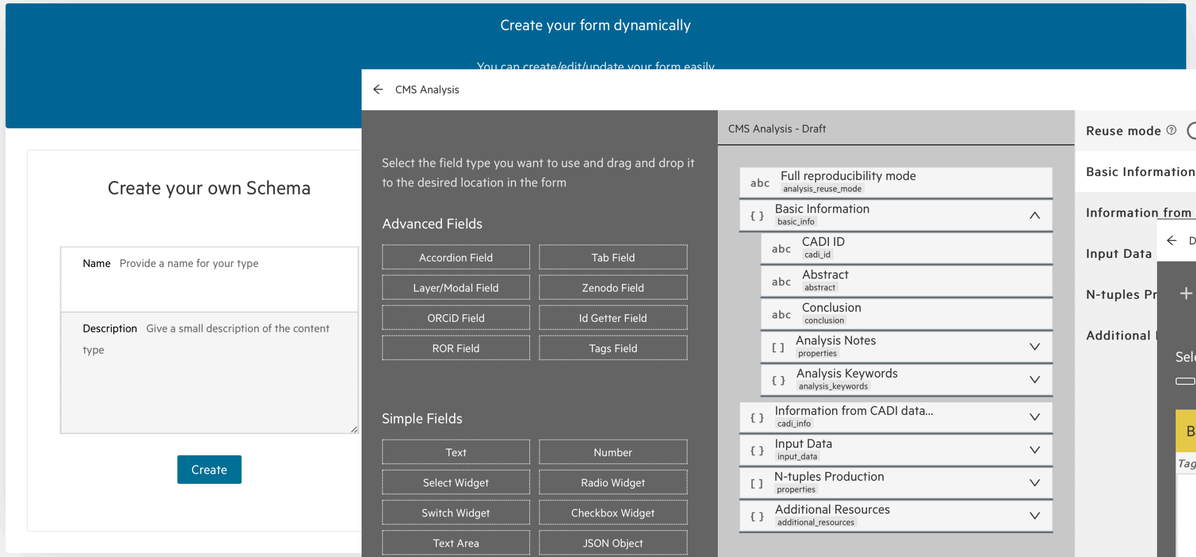
```
$ export CAP_SERVER_URL=https://analysispreservation.cern.ch/
$ export CAP_ACCESS_TOKEN=<generated access token from server>
```

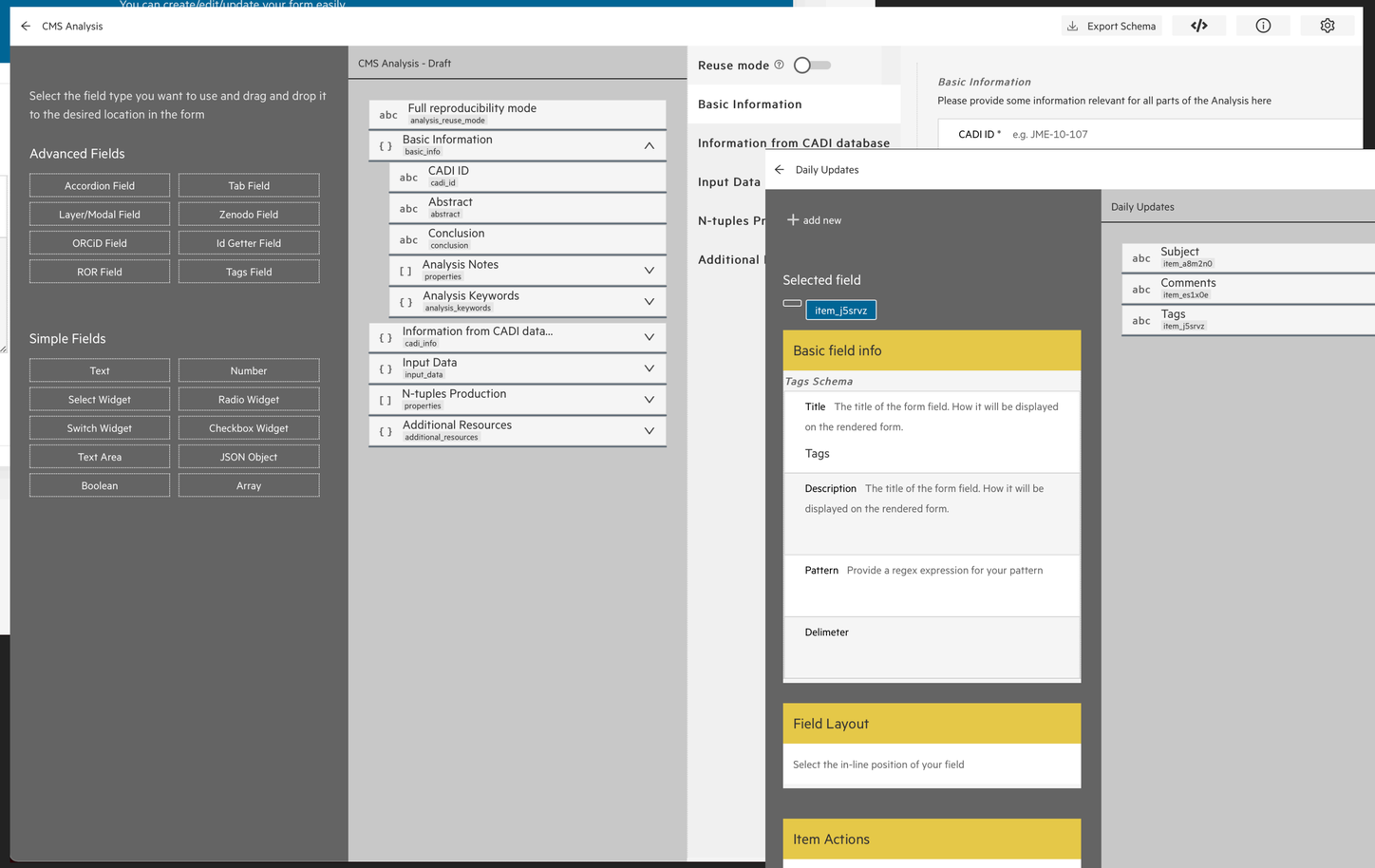
```
$ cap-client me

{
   "collaborations": [
      "ATLAS",
    ],
   "id": 1,
   "email": "atlas-user@cern.ch"
}
```

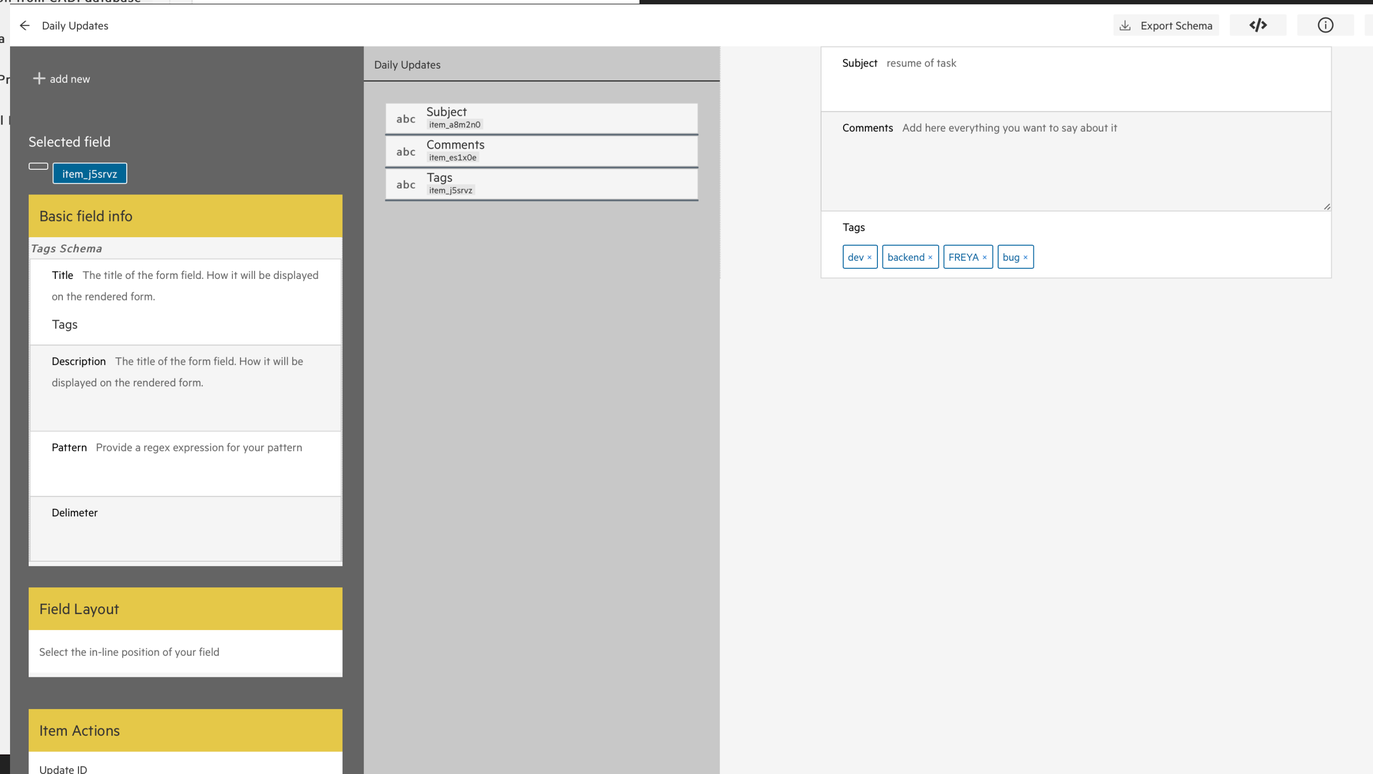
```
$ cap-client files list --pid/-p <existing pid>
 [
   {
      "checksum": "md5:f0428126e7cf7b0d4af7091c68ae2a9f",
      "filename": "file.json",
      "filesize": 25,
      "id": "25852e50-be6d-47a5-897b-1f3df015fac7"
   },
   {
      "checksum": "md5:926fb9c44251d70614ee42d34c5365b6",
      "filename": "Analysis_Notes_07112019.pdf",
      "filesize": 160898,
      "id": "89743c9b-106d-4235-8e96-23a164c7b1f4"
   }
 ]
```

FORM/CONTENT BUILDER

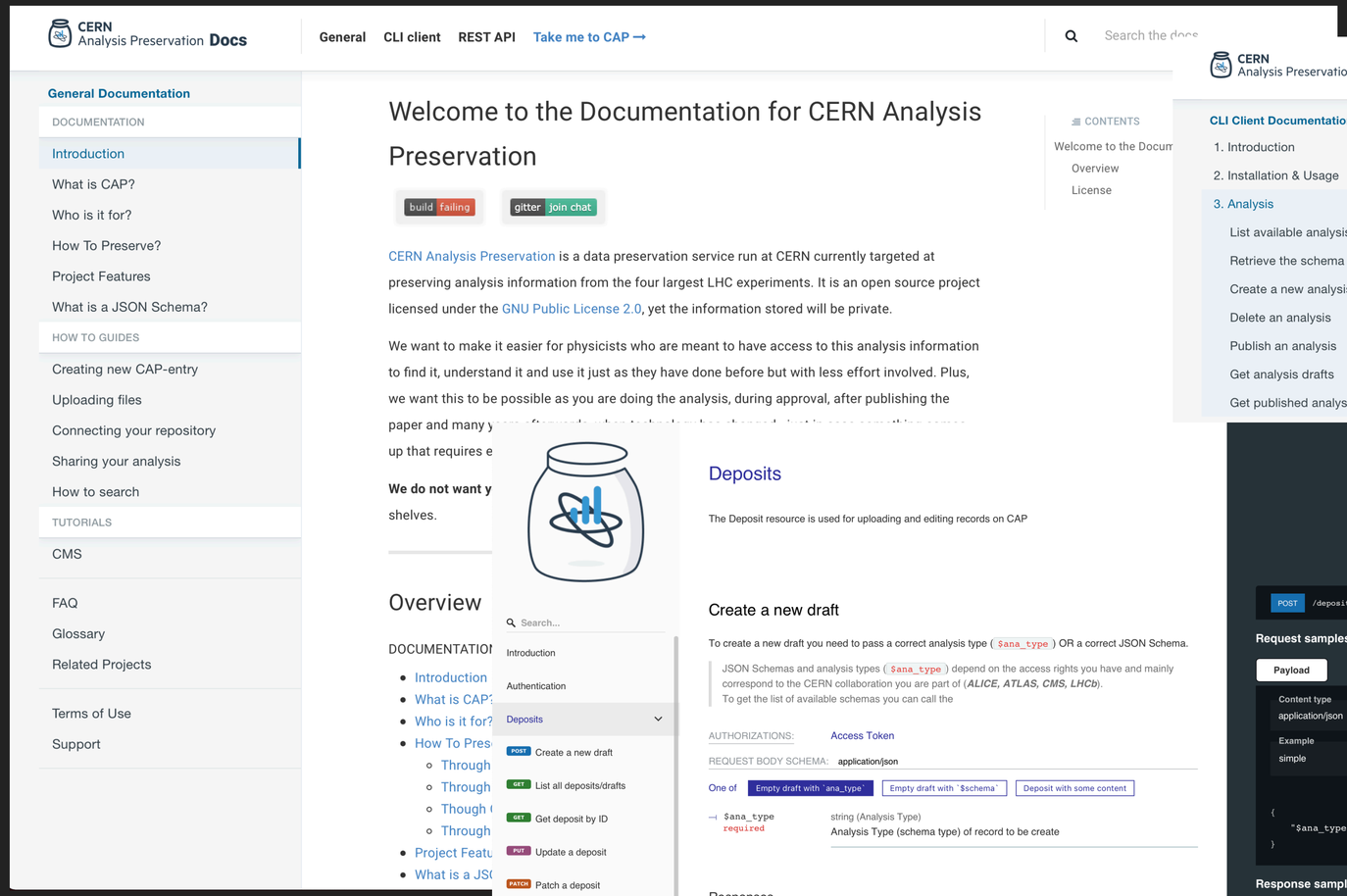
"Admin" interface for customising forms/search/experiment spaces

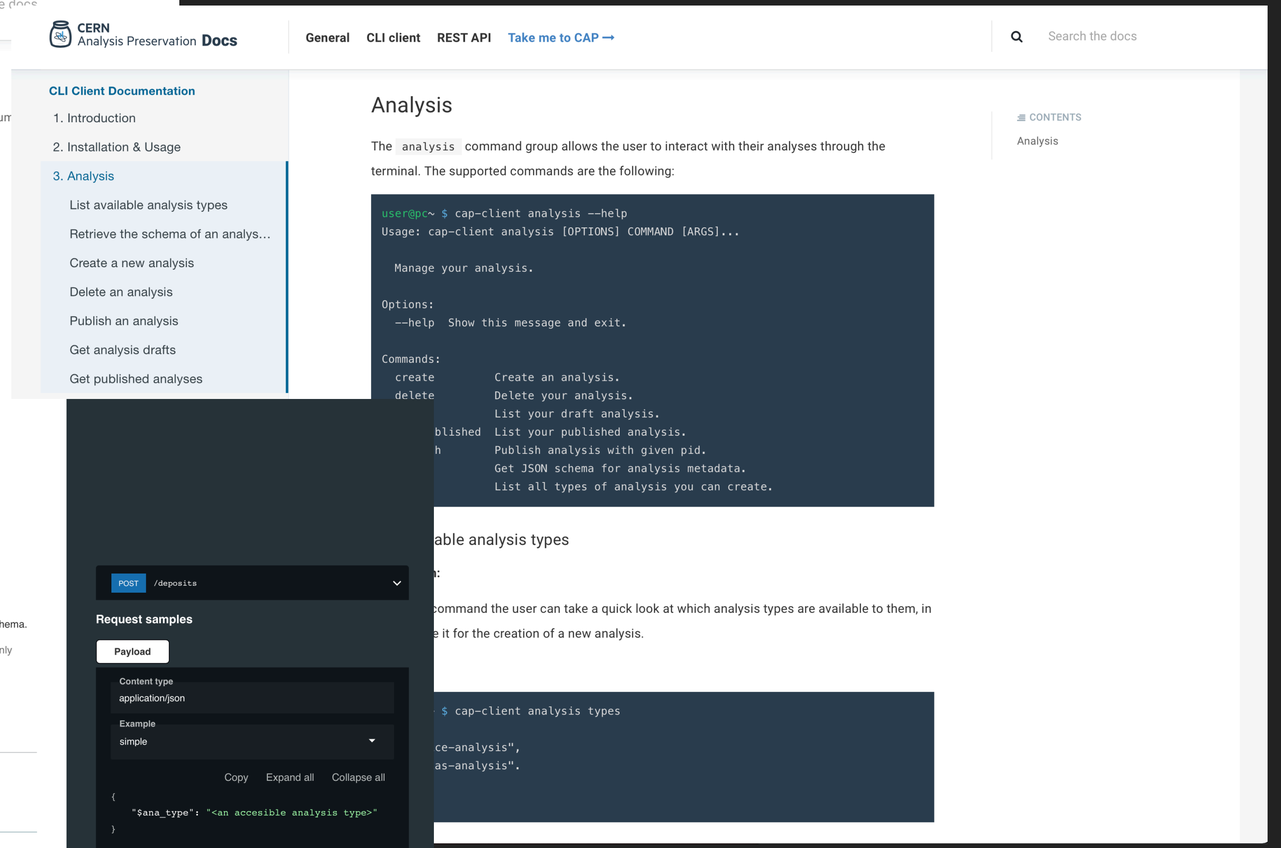


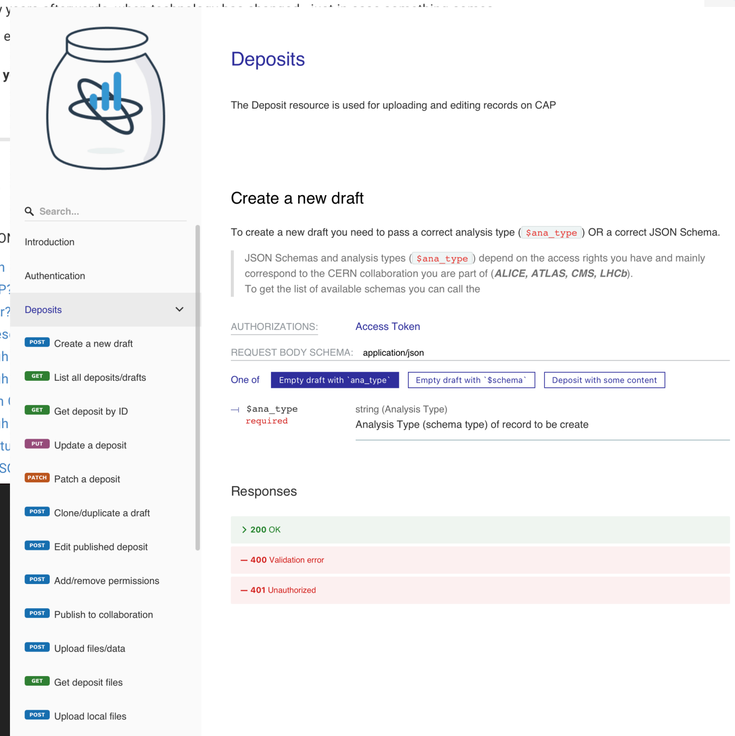


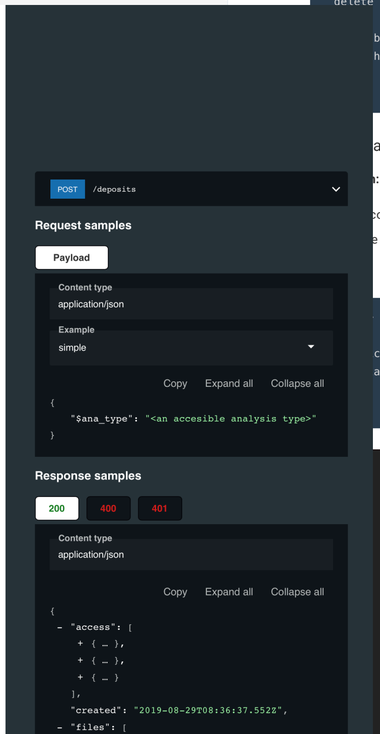


DOCUMENTATION









CONTENT

- Customisable to many needs
  - metadata, metadata & files, metadata & files & workflows
  - reviewable
  - Repository connection, syncing
  - other combinations of above
- Multiple versioned models
- Searchable
- Easily accessible (browser, REST API, CLI client)
  - Useful for people, BUT also machines
- GUI Admin area for data modelling (coming soon!)

CONTENT: CMS COLLABORATION

With the help of the CAP platform, CMS is building a nice collaboration space

CMS Analysis - example

- Normal/Reuse mode
- CADI sync/querying
- DAS import/sync/querying
  - Same query language as DAS-client
  - Faster search!
- Triggers
- CADI Analysis Keywords
  - Harvested from Google Spreadsheet
  - Made searchable
  - Connect with analysis

** CMS team has started manually populating analysis harvested from CADI. (Next: more content, better search capabilities, more feedback to come)

CMS Statistic Questionnaire - example

- Mandatory step on the CMS Analysis “process”
- Approval workflow
  - Stats committee reviews
  - Documents reviewable by specific e-group, users
- Legacy migration (old website migrated to CAP)
- Email notification on actions (publish/review/etc) e.g.:
  - (1) normal email with info
  - (2) "trigger” email to tools like hyper-news, Jira, etc
- Some improvements from legacy
  - Possible and easier to link with analysis description
  - All info stored in one place
  - Searchable

CMS Muon Group - coming soon

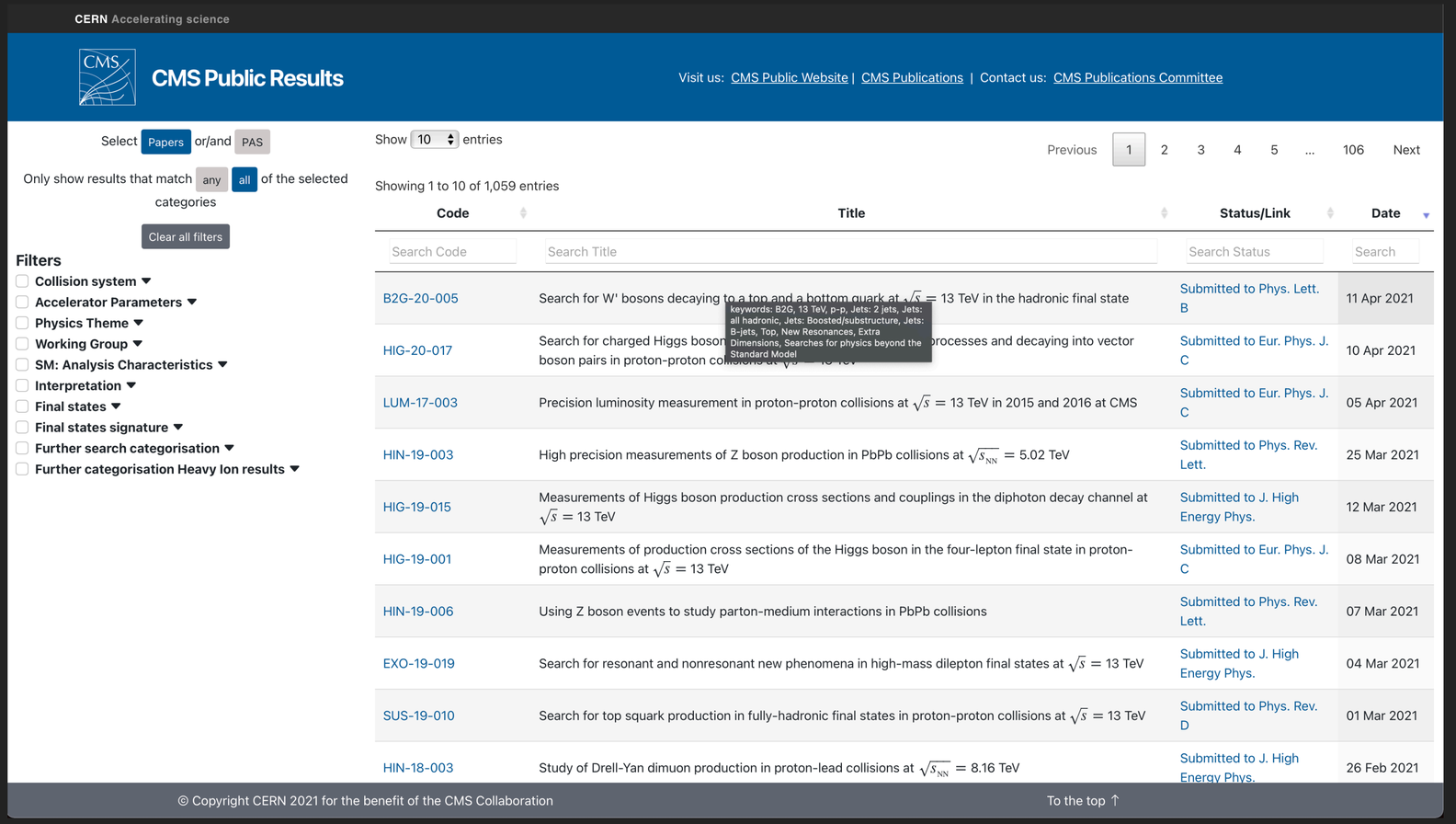
- Currently exist in Twiki pages
  - Not searchable
  - Can’t be linked
- Make documents reviewable from specific people

USE CASE: CMS COLLABORATION

Reuse of CAP data

CMS public results page

- Fetching data from CAP
- Dedicated website
- Example of reuse of information



OTHERS ARE FOLLOWING..

ATLAS - preserving the full analysis cycle

- a Glance Record
- a number of repositories
- a number of docker images
- EOS location with static assets needed for RECAST repository with the REANA-ready workflow space
- Possibility for a preserved likelihood data product

LHCb

- Currently "minimal analysis preservation" requirements at LHCb
- Plan to refactor the analysis procedure, with preservation in mind
  - Looking for better DB solution for staged submission-forms (workflow model)
  - Different roles => different responsibilities/actions
  - Data reuse: Trigger continuous building of websites (WG website, publications website, etc.)

FUTURE

Goal: Develop working use-cases in all 4 LHC experiments

Goal: Make CAP the general tool for all collaboration

Scalability: move to Kubernetes!

Platform:

- Complete development of REANA integration
- Release “Content Builder” feature to admins
- Do more user testing and expand the user base
- Collection pages: dedicated page displaying published results, guides, access rights, etc
- EOS integration for user based file storage/retrieval
- SIP integration with Archivematica instance at CERN (2nd level of preservation)

Possibilities: Transform CAP to a domain agnostic knowledge and data preservation platform!

PORTALS, SERVICES AND TOOLS

CERN REPRODUCIBLE ANALYSIS FRAMEWORK

Web: https://analysispreservation.cern.ch
Github: https://github.com/cernanalysispreservation
Email: analysis-preservation-support@cern.ch

Web: http://reanahub.io
Github: http://github.com/reanahub
Email: info@reanahub.io
Twitter: reanahub

Web: https://opendata.cern.ch
Github: https://github.com/cernopendata
Email: opendata-support@cern.ch

THANK YOU