CERN Analysis Preservation Documentation

CERN Analysis Preservation Support Group

Dec 05, 2018

Contents

We want to make it easier for physicists who are meant to have access to this analysis information to find it, understand it and use it just as they have done before but with less effort involved. Plus, we want this to be possible as you are doing the analysis, during approval, after publishing the paper and many years afterwards, when technology has changed - just in case something comes up that requires expertise for your analysis. In short:

We do not want your work to get lost. We would rather like to give it a very special place in our shelves.

CHAPTER 1

Further Reading

1.1 Introduction

CERN Analysis Preservation (CAP) is a service to describe, capture and reuse analysis information in HEP (High Energy Physics). It is developed by the collaborative effort of information and computer scientists at CERN, in partnership with the LHC collaborations.

1.1.1 Reproducible and Reusable Research

The initial idea behind the project was to preserve analyses for the purpose of reproducible research, making it accessible, understandable and reusable in many years to come. In conversations with LHC physicists, it became apparent that the information we collect will be valuable not only in the future, but already from the very start of information taking.

1.1.2 Analysis Definition

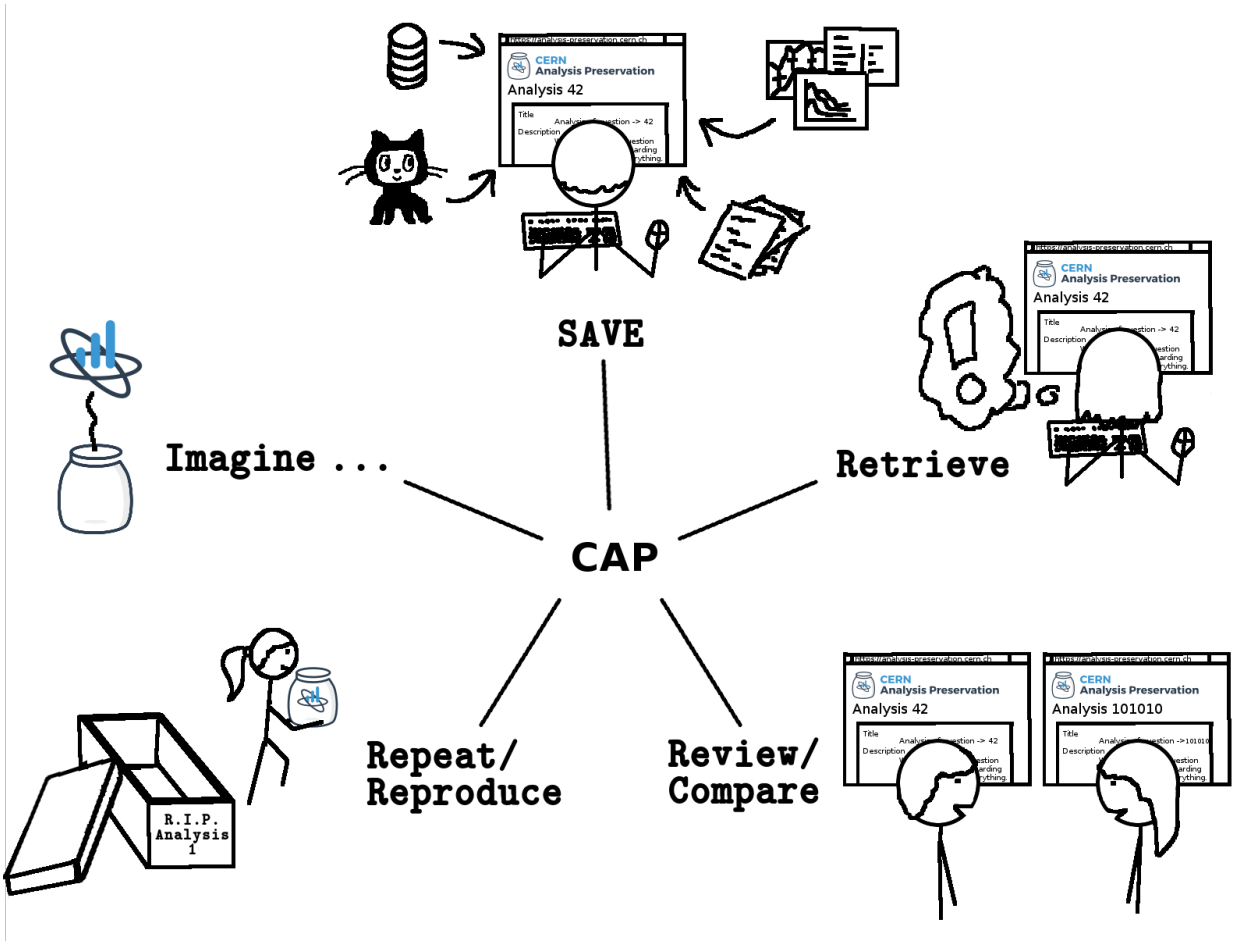
To us, an analysis consists of both data (e.g. datasets, code, results) and metadata (e.g. analysis name, contact persons, publication) the combination of which we call analysis information. While we structure this information on CAP in a way that represents the analysis workflow steps, we do not require or ask for any change in the physicists' individual workflow or the terminologies used in different collaborations and working groups. To accommodate for any changes in content and workflow of an analysis, we keep versions of both the analysis record itself and the underlying JSON schema (for more information see *JSON Schema*).

1.1.3 Access Control

As we are preserving sensitive data, we take care to apply safety measures and access control to any information added to CAP. Access will always be restricted to members of the collaboration associated with the analysis. Permissions within a collaboration can be adjusted by the creator of the analysis, defaulting to creator-only access. For more information please refer to our section on *Authorisation and Access Control*.

1.2 Project Features

CERN Analysis Preservation is a centralised platform for physicists to preserve and document information relevant to their analysis so that it remains understandable and reusable in the future.



- *1. Preserve*
- *2. Search and Retrieve*
- *3. Review and compare*
- *4. Reuse*
- *5. Imagine...*

1.2.1 1. Preserve

We try to make it as easy as possible for you to preserve your analysis information. To achieve this, we have implemented a web form and an API including a dedicated client, so that you can submit or update content from your shell. In addition, we have established connections to collaboration databases fetching as much information as possible automatically and offering different possibilities for adding the rest.

You will be able to start adding information at any time, be it at the very beginning of your analysis, in preparation for a conference, at the time of publication or at any time it is useful.

This section will give you some background knowledge on how we save, describe and capture analysis information and who is able to access it.

Submission Form

The submission forms are there to help you submit the different materials you create(d) while doing your analysis. They are a graphical representation of the JSON schemas. Information stored within connected databases is used to auto-complete and auto-fill the analysis description whenever possible. See the section on *Connections to other databases* for more information. There are tutorials to help you *create an analysis* submitting your analysis details. While we have taken steps to sure that the JSON schemas work for most people in a collaboration, we acknowledge that there may be differences in the way different people conduct an analysis. If you don't see that your analysis "matches" the schema provided, please get in touch with us. The forms also allow manual editing of each field, as well as the ability to input JSON directly.

Submitting via the shell - Our "CAP-client"

Using your shell, you can use the CAP-client to submit, update, retrieve an analysis and its components without the need to use the web forms. You can find details on how to install and use the client in http://cap-client-test.readthedocs.io/en/client-docs/#

REST API

There is an API to enable direct interaction with the content, provided you have the right access permission. Help on using the API will be provided *here.*.

Connections to other databases

To save time when submitting and to ensure accuracy of added information, we are connecting to collaboration databases and systems containing analysis information. This allows us to auto-complete and auto-fill most of an analysis record as soon as it is created, given that the content exists in the databases. GitLab integration is also in place, so that you can automatically fetch analysis details, e.g. images, from there.

Additionally, we offer the possibility to uploading files (e.g. configuration files), as well as providing a URL, from which the files are copied and stored automatically.

For more details on how these integrations work, you can go to the *tutorial section*.

Versioning

Upon creation of an analysis on CAP, a unique identifier is assigned. Every time the analysis is edited, the new version will be stored as an update to the previous version of the analysis through the identifier system. This will enable references to intermediate analysis steps in the analysis notes and allows keeping track of the analysis.

Authorisation and Access Control

Authorization on CAP is managed by CERN Single Sign-On, therefore applying the usual access restrictions you are used to from your collaboration.

Due to the sensitive nature of analysis information and content - especially in early stages of an analysis - accessibility of analysis information is subject to permissions set by the collaborations, as well as the creator of an analysis and the collaborators involved.

When starting a new analysis submission to CAP, the analysis record is saved as a draft. By default, the creator of the draft record will be the only one able to view and edit it. Read or edit rights can be granted to researchers in the analysis team or the working group.

As soon as the analysis is "published", the analysis will be shared with the collaboration, meaning its members will acquire rights to view the analysis. Editing rights will remain as they were for the draft. A draft version can be submitted any time. We encourage you to deposit the analysis as soon as possible so that it becomes "visible" to the members of your collaboration (and no one else). However, these decisions are up to you and the collaboration's practices.

Note:

- only collaboration members have access to a collaboration's area, can create analyses and can see shared analyses
- only a certain collaboration's members have access to this collaboration's analyses
- only members granted specific rights can see or edit a draft version of an analysis
- only the creator can see or edit an analysis with default permission settings

---

1.2.2 2. Search and Retrieve

The search capability of CAP can help users find both preserved and on-going analyses they have access to in CERN Analysis Preservation.

Search capability

Using the search bar at the top of the page or the dedicated search page that comes with it, users can search through their own and all shared analyses within their collaboration, past or on-going. Filters (=facets) will help you select the relevant content. All analysis metadata are indexed, which means users can find analyses with specific parameters, processed with a specific algorithm, or using a specific dataset or simulation to name a few examples. Information that is not explicitly added to the schema and instead stored in an uploaded file are not indexed for search right now.

Note: You have suggestions on what is needed to make the search more useful to you? Please *let us know*!

1.2.3 3. Review and compare

CAP aims to support reviewing analyses and with that the process of analysis approval by enabling the user to give specific access to analysis records and store relevant analysis information in one place. If the collaboration decides so, relevant information could be exported easily to tools like Indico, for example. Exporting a record is liable to the same restrictions as accessing the record.

1.2.4 4. Reuse

In CAP analyses information is preserved with the aim of reusing it - now or in the long term. We are working on making that easy as well! In the REANA project we build a framework to enable easy reinstantiation of an analysis. See *this list* for a short description of these related projects.

1.2.5 5. Imagine...

The above use cases were derived from input we received from CERN physicists. We are open to new ideas, which is why everything you want to do with your analysis information that will help you with your research is part of what describes CAP.

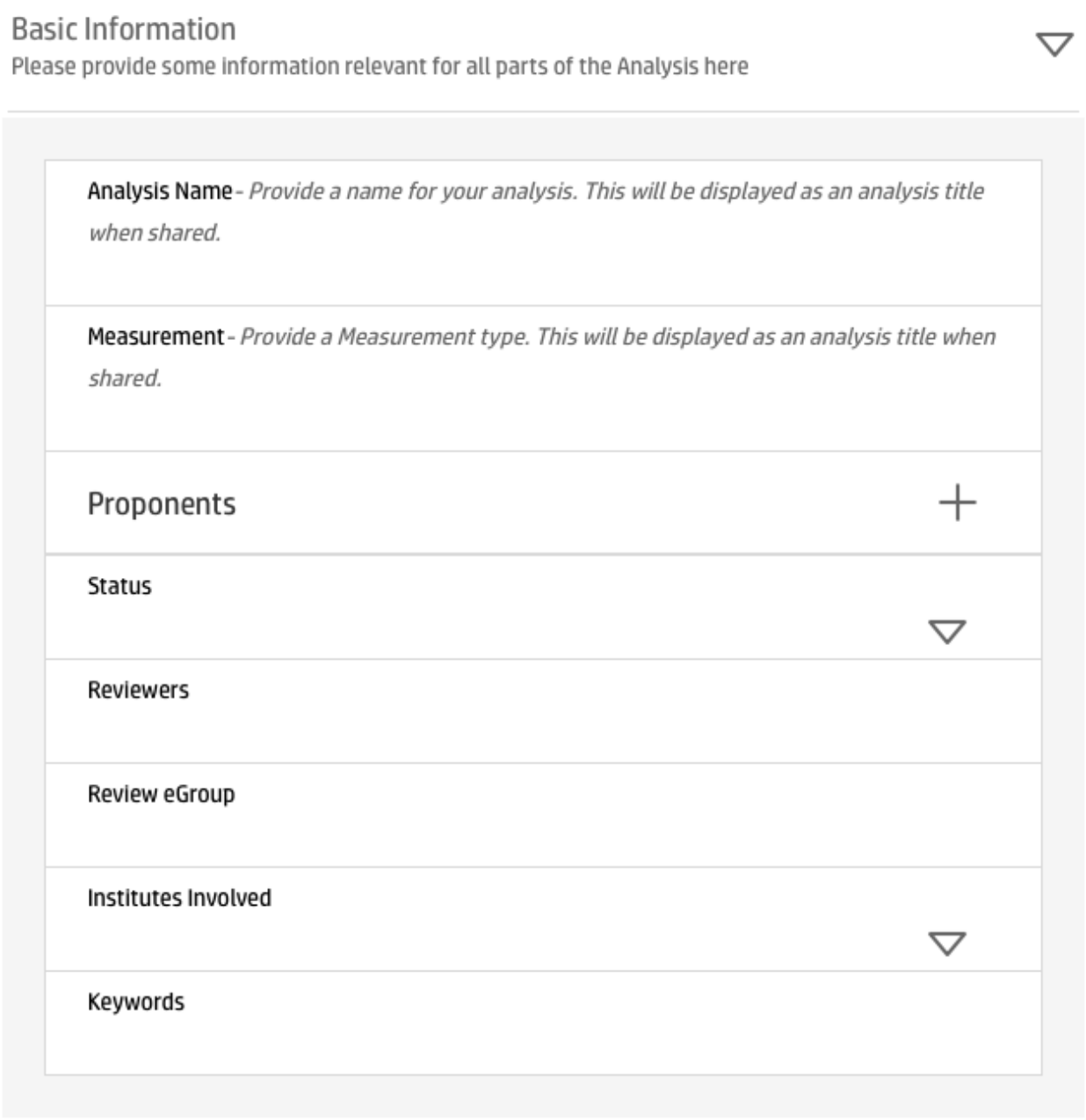
1.3 Tutorials

1.3.1 The CAP form

There are four main sections to document a physics analysis on the CAP form: basic information, data provenance, analysis software and documentation. However, each form is tailored to the needs of each experiment. In the following examples, we use form snippets from different experiments to demonstrate their flexibility and use.

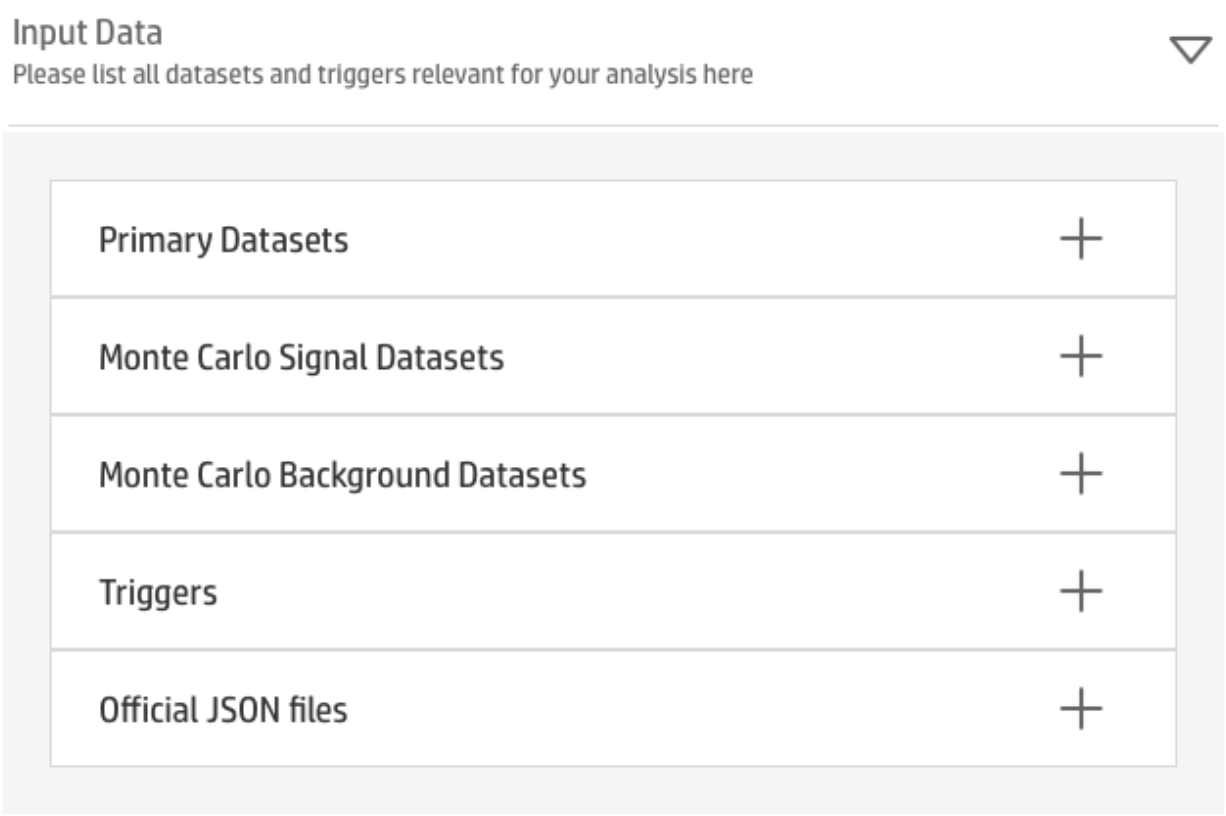
Basic information

Basic information captures name, measurement, proponents, the status of the application and other.

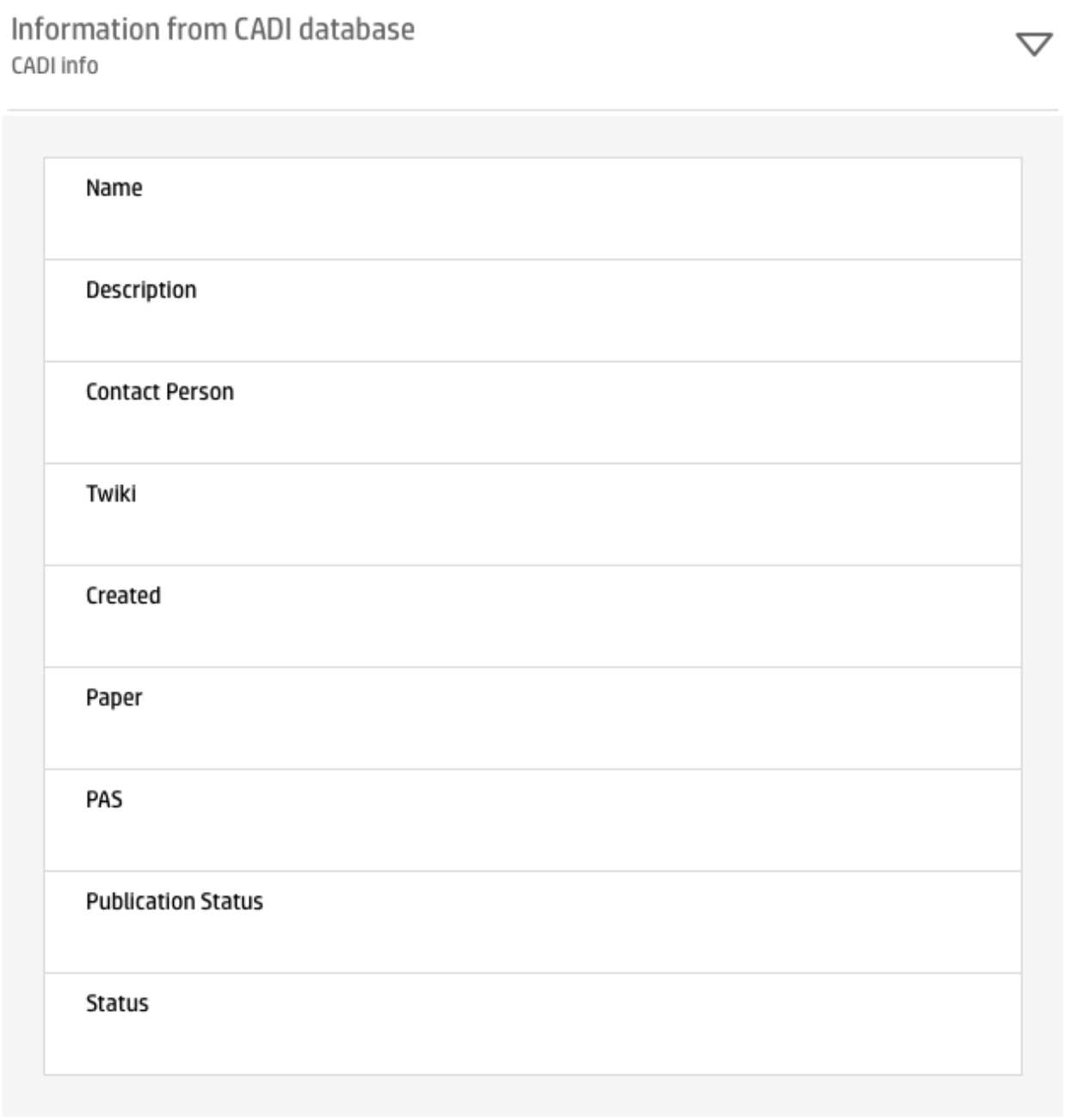


Data provenance

The section on data provenance captures which data sets are used in the analysis and how. By clicking the button **+** a large number of data sets can be documented.

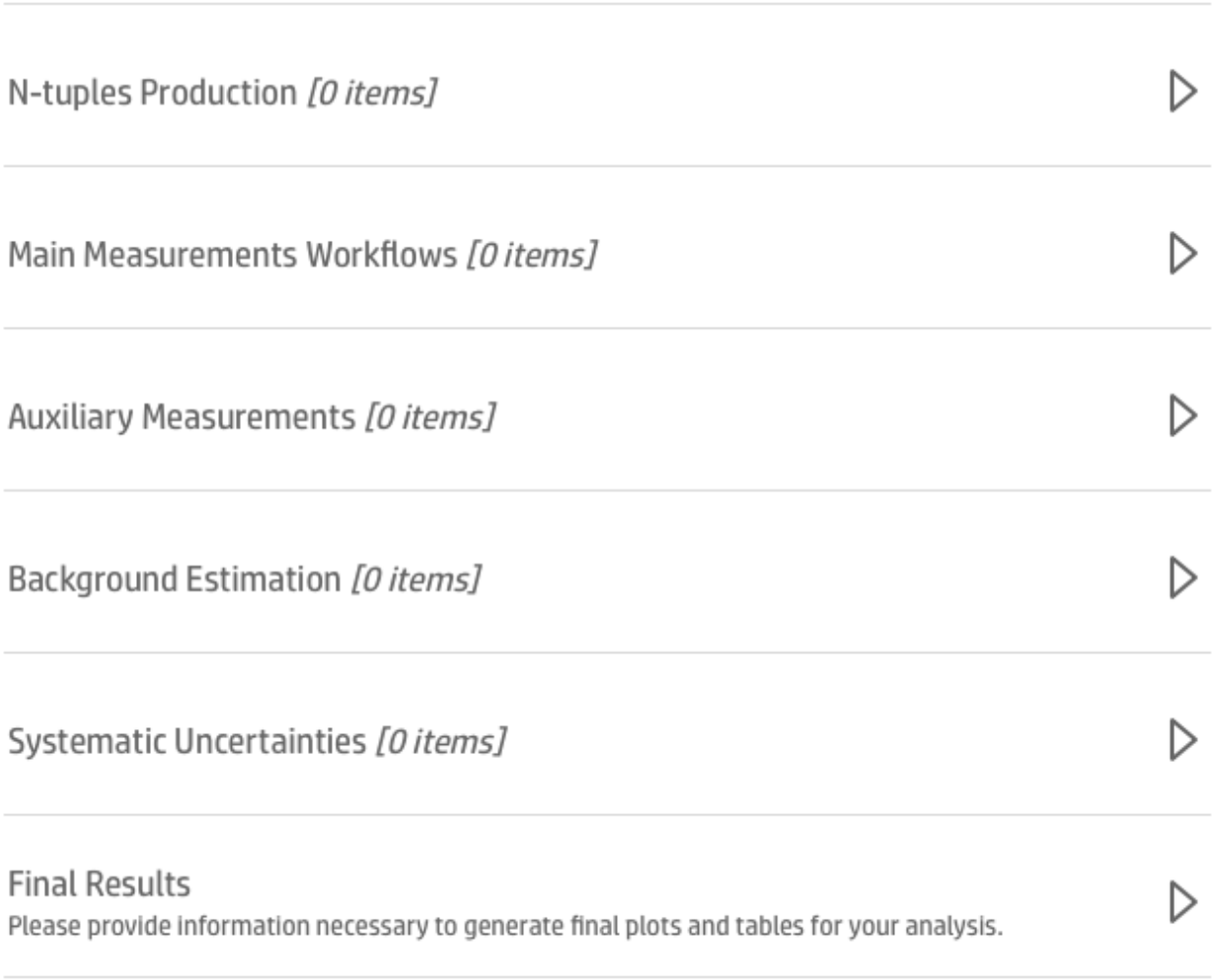


The CAP form implements the links to the existing experimental databases. In the example below an analyst can import their analysis information from the CADI database at CMS.



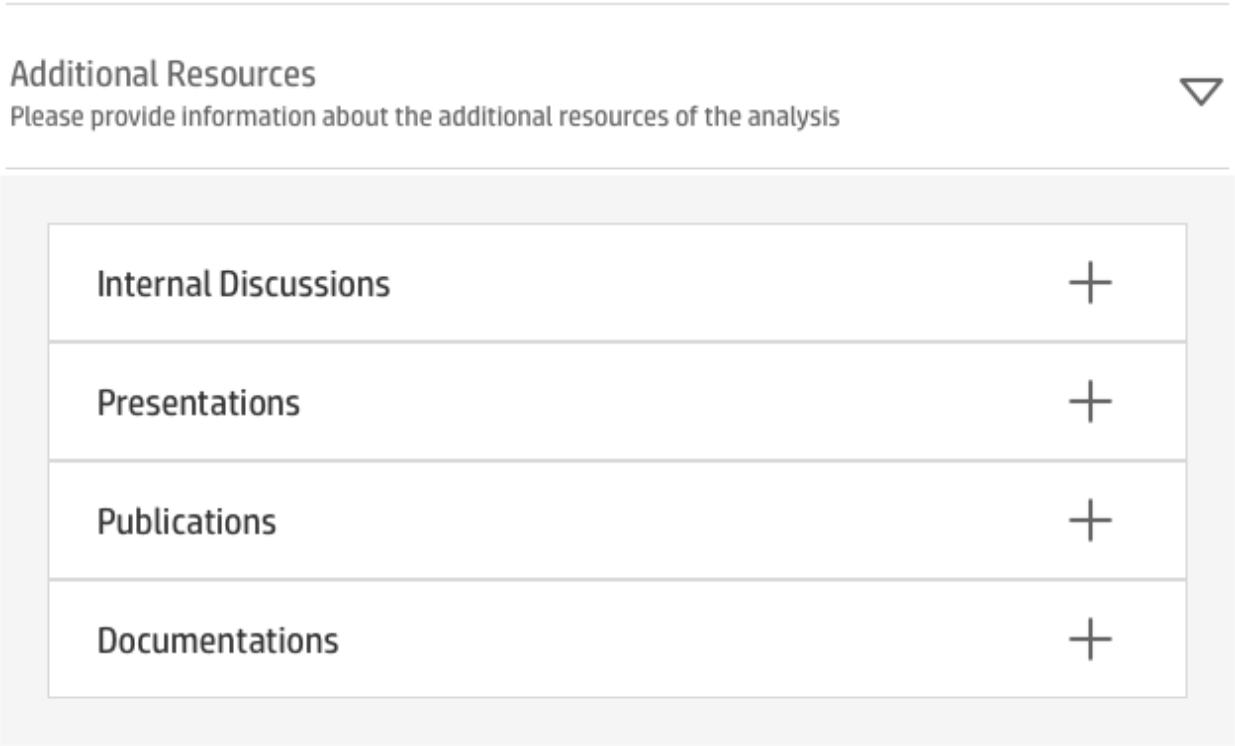
Analysis software

Analysis software can be captured directly from git repositories. Analysis workflows which facilitate analysis automation can also be documented with the form.



Additional resources

The additional resources section captures presentations, publications and other internal documentation.



1.3.2 CAP-client

The CAP-client is a command-line tool for preserving analyses. It is implemented as a python package and its documentation can be found in the CAP Client docs.

Setting up the cap-client configuration:

The basic communication with the server can be seen here:

Editing analysis metadata:

1.3.3 Reusable Analyses REANA

REANA is a reusable and reproducible research data analysis platform. It helps researchers to structure their input data, analysis code, containerised environments and computational workflows so that the analysis can be instantiated and run on remote compute clouds. REANA was born to target the use case of particle physics analyses, but is applicable to any scientific discipline. The system paves the way towards reusing and reinterpreting preserved data analyses even several years after the original publication. Find comprehensive documentation about the REANA project.

1.4 JSON and JSON Schema

The analysis information is modelled in JSON to ensure data is added in the structure and formatting predefined by the CERN Analysis Preservation team in and the physicists from the LHC collaborations. JSON is an open data format

where data is represented as objects of key-value pairs. It is independent from any tools and programming languages, usually stored in .json text files and can be parsed and created to and from most programming languages and systems. Therefore it is an easy format to share and preserve data.

An example miniature JSON file:

```
{
  "basic_info": {
    "analysis_name": "Z -> ee",
    "status": "0 - planned / open topic",
    "people_info": [
      {
        "name": "John Doe",
        "email": "john.doe@cern.ch"
      },
      {
        "name": "Jane Doe",
        "email": "jane.doe@cern.ch"
      }
    ]
  }
}
```

The above definitions contain a name, status and an array of names and emails. Every `{...}` marks an object, `[...]` marks an array and strings are wrapped in `"..."`. Other possible data types are numbers (e.g. `1`, `0.5`), booleans (`true`, `false`) and `null` for an empty value.

A JSON schema defines a structure and rules that apply to it. JSON data can be validated against a schema to check whether the data fits the pre-defined structure and requirements of the schema.

An example schema against which the above JSON file is validated:

```
{
  "$schema": "http://json-schema.org/draft-04/schema#",
  "title": "Internal Title (not displayed)",
  "description": "Internal Description (not displayed)",
  "type": "object",
  "properties": {
    "basic_info": {
      "title": "Basic Information",
      "description": "Please provide meta-data information relevant for the analysis␣
↪here.",
      "type": "object",
      "properties": {
        "analysis_name": {
          "title": "Analysis Name",
          "type": "string",
          "required": true
        },
        "status": {
        "title": "Status",
        "type": "string",
          "enum": [
            "0 - planned / open topic",
            "1 - in preparation",
            "2 - ANA note released",
            "3 - review committee",
            "4 - collaboration review",
            "5 -",
```

(continues on next page)

(continued from previous page)

```
            "6 - CONF note published",
            "7 -",
            "8 - journal review",
            "9 - PAPER published",
            "x - other"
          ]
        },
        "people_info": {
          "title": "Proponents",
          "description": "Please provide information about the people involved in the
↪analysis.",
          "type": "array",
          "items": {
            "type": "object",
            "properties": {
              "name": {
                "title": "Name",
                "type": "string"
              },
              "email": {
                "title": "Email-Adress",
                "type": "string"
              }
            }
          }
        }
      }
    }
  }
}
```

The schema defines the structure and data types JSON data needs to follow and use to validate against it and allows to add additional rules like required fields.

Each schema is directly provided or created with the support of collaboration physicists and tested, as well as revised, several times to ensure that important information will be preserved. Throughout this process, core components of an analysis are identified and structured. Each collaboration has its own unique schema to capture the workflow that fits their specific requirements.

Every schema change is versioned so that it can adapt to changes in the data or other components provided by the collaborations. This practice also ensures the that the integrity of the older analysis records is maintained.

Depending on the preference and work environment of the researcher, analysis information can be created and edited through a *Submission Form* on the web interface or via the *Submitting via the shell - Our "CAP-client"*.

1.5 API Reference

CERN Analysis Preservation offers a REST API to access the service independently from the web interface. If you want to automate specific tasks or create your own data interface, you can use the API to do so.

1.5.1 Aquiring an Access Token

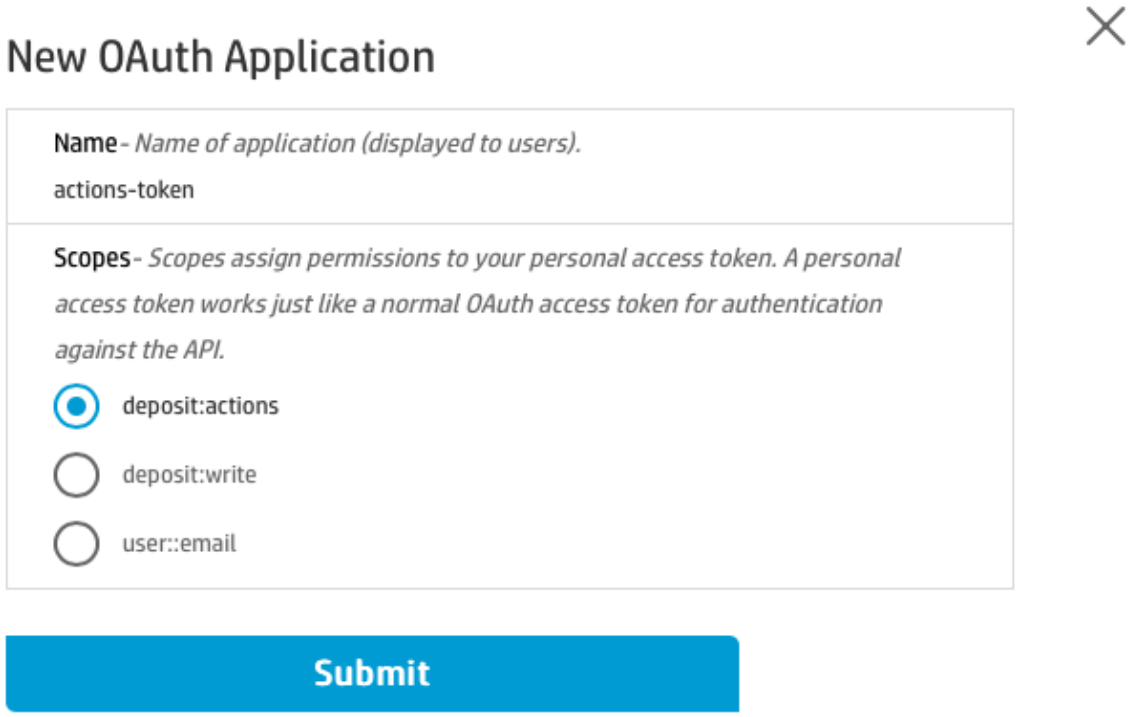
If you want to gain access to CERN Analysis Preservation from your console or any external means other than the web portal you will need an access token to authenticate with the portal. You can create multiple tokens for different

services.

Warning: Your access token will allow you to use the service in the same way in which you may use it if you log in on the web portal. You will have the same permissions unless specified otherwise on creation of the token. This implies that anyone who has this token can log in as yourself to the service. Do not share your personal access token with anyone else, and only use it with HTTPS!

To get an access token, you will need to log in on the web portal and create one, as shown below.

In this dialog, *scopes* lets you define permissions for the token which by default only include read access to your drafts and records.



Clicking *submit* will generate and show your personal token in the browser. Please copy it to a safe place on your computer, as it is not stored on the portal and you will not be able to retrieve the same token again in the future.

1.5.2 Accessing the API

Access your drafts using your token in the following link:

```
https://analysispreservation.cern.ch/api/deposits?access_token=TOKEN
```

and your shared records with the following link:

```
https://analysispreservation.cern.ch/api/records?access_token=TOKEN.
```

Adding the ID of a specific record or deposit in the link will give you access to this particular one only:

```
https://analysispreservation.cern.ch/api/deposits/<id>?access_token=TOKEN
https://analysispreservation.cern.ch/api/records/<id>?access_token=TOKEN
```

1.6 Terms of Use

CERN Analysis Preservation facilitates the collaborative nature of the LHC collaborations. The access restrictions are those selected by the respective LHC collaboration. Do not share materials outside the foreseen access restrictions or otherwise circumvent CAP's access restrictions. Many materials that can be found through the service might be work in progress, so they should be used with caution. According to collaboration practices, it is not allowed to publish materials openly that are not approved by the respective committees within the collaboration.

CERN Analysis Preservation is not liable for any content on the portal. It is the responsibility of the individual researchers and working group to preserve, update and share content in a timely manner.

Furthermore, the use of the CERN Analysis Preservation service denotes agreement with the following terms:

- The service is provided free of charge for the individual user;
- Download and use of content from the service does not, unless expressly stated in the applicable license conditions, transfer any intellectual property;
- Bulk downloading of personal data taken from the service is not allowed;
- All content is provided "as is" and the user shall hold the service and the content providers free and harmless in connection with their use of such content; and
- The service providers reserve the right, without notice, at their sole discretion and without liability, to restrict or remove user access where they consider that use of the service interferes with its operations or violates these Terms of Use or applicable laws.

If you encounter and problems or have any questions, you can contact us at analysis-preservation-support@cern.ch.

1.7 Frequently Asked Questions

Here you can see frequently asked questions:

1.7.1 Why should I use CERN Analysis Preservation?

We believe that your research is worth preserving. CAP is there to support you in making sure that the work and thoughts you have put into your analysis last beyond the publications. By using CAP, you are able to safeguard your analysis resources (data, code, containers, etc.) and move on to the next project. In addition, you can easily search through other submitted analyses (details) that you find interesting. With the REANA project, you will also be able to directly rerun these analyses. We know it is hard to fully document physics analyses, so we try to make it as easy as possible for you to submit and update content.

1.7.2 How do you define an "analysis"?

A physics analysis usually comprises numerous files, from data, code, workflows, containers to various configuration files. Also, often there is a lot of contextual information ("meta-information") that went into the analysis, which is essential for understanding the analysis in the future. From our point of view, an analysis is a combination of data and metadata. However, it should be noted that we use the term "data"

very loosely. While every analysis is different, many core components are the same, which is apparent in the JSON schemas we provide. For more details, please see the *Analysis Definition*.

1.7.3 Is the information or data I add openly accessible?

No. You assign the access permissions to your analysis. Access can be restricted to your collaboration or according to your preferences (see the section on *Authorisation and Access Control* for more details). Nothing is by default open access on CERN Analysis Preservation (except for the projects' own source code). It is designed to be a safe environment for CERN physicists to use from the very beginning of starting their analysis and at any given moment in its lifetime.

If you are searching for a service providing open access data or you have some data to share, you may want to check out http://opendata.cern.ch/.

1.7.4 I can edit, but can my collaborators edit my analysis too?

This depends on what permissions you assign to your collaborators. All your collaboration colleagues should be able to read it (read-only access), but only those you invite specifically (by email or e-group) can edit as well.

1.7.5 Are you trying to automate all analyses?

No. Physics analyses are incredibly complex research work, which is precisely the reason why they are invaluable for preservation. What we are trying to do is help you with repetitive tasks, help you find the information you need, and support your review and approval process so that you can focus on the actual research.

1.7.6 Do I have to use this rather long form or what other options do I have?

You can use your shell and our "CAP-client" to submit, update or find an analysis. That way, you can entirely circumvent the long submission form. Also, by using collaborative tools like GitLab, submitting your analysis becomes much more comfortable. You can automatically connect your repository to CAP and preserve different versions of the code. For information on how to use the CAP client, please consult the client's documentation: https://cap-client.readthedocs.io/en/latest/

1.7.7 What is the ID for the analysis?

We use a unique ID for your analysis to distinguish it from the other analyses available on CAP. You can use it to update your analysis or find information about it via the CAP client. Once you submit your analysis, you will see the ID in the URL of your analysis page on CAP.

1.7.8 As a database provider within LHC collaborations, how can I contribute to or profit from CAP?

There are many ways you can support us and we can support you. Please contact us at analysis-preservation-support@cern.ch to find out more about our current efforts.

1.7.9 I have no idea what I am doing here, can you help me?

Sure, contact us by email at analysis-preservation-support@cern.ch.

1.8 Glossary

Here is a list of terms we use regularly, so you get an idea of what they are:

Deposit A draft analysis decription on CERN Analysis Preservation.

Docker/Docker Container An open source software that allows arbitrary software to be wrapped inside a so called "Docker container". This container mimics the environment the software usually runs in. Thus, it can be preserved and run relatively easy.

Invenio-Deposit The part of the *Invenio* framework that in case of CERN Analysis Preservation handles everything directly related to analysis records like permission to view and edit and storage.

Kubernetes An open source software used by *REANA* for managing containers (usually Docker containers) on a cluster. This includes scheduling and scaling tasks and executing the software wrapped inside the containers.

UMBRELLA A container tool similar to Docker. It is designed as a light-weight tool for preserving an environment while considering hardware, operating system, software and data.

Yadage Workflow A set of JSON schemas with rules how to chain them together to describe analysis workflows. They are wrapped in a container (e.g. Docker) and can be executed by Yadage both locally (e.g. on your laptop) and distributed (e.g. on Kubernetes or *REANA*).

1.9 Contact Us

If you need help or have a question, please contact us using our Gitter chatroom or email us via analysis-preservation-support@cern.ch.

In case direct contact is not what you are looking for, you can comment on or add to the issues on GitHub.

1.9.1 Team and Community

Our core team is based at CERN, in CERN IT and in the Scientific Information Service. We are computer engineers and information scientists, and we work closely together with preservation and outreach representatives from the LHC collaborations as well as related projects like REANA, RECAST, DPHEP. You can find brief descriptions on these projects in *this list*.

Many contributions are made by or facilitated through the help and support from collaborators from the LHC collaborations.

Contact us if you want to contribute or have questions!

1.10 Related Projects

CERN Analysis Preservation is related to a couple of projects, some of which may come up in discussions and are therefore listed here:

CAP An acronym for CERN Analysis Preservation. If you do not know what that is yet, take a look *here*.

COD/CODP Acronyms for the CERN Open Data Portal. It is an open-access portal for CERN experiment data and software and serves as a learning platform as well as enabling further research and exploration.

DASPOS An acronym for Data and Software Preservation for Open Science. A project to explore preservation possibilities and techniques in high energy physics.

DPHEP An acronym for Data Preservation in High Energy Physics. A study group to explore preservation possibilities and techniques in high energy physics.

HEPData An open-access portal preserving and providing data, plots and tables from publications in high energy physics.

Invenio An open source digital library framework developed at CERN that CERN Analysis Preservation is based on. It provides background functionality like authorization, working with analysis records and storage.

REANA An acronym for Reusable Analysis. A system that schedules and runs analyses on the CERN cloud based on *Kubernetes* and *Yadage Workflows*. It is used to rerun analyses from CERN Analysis Preservation and RECAST.

RECAST A service based on requests to re-execute an analysis chain with the possibility of using a new signal model. Analysis chains are defined and stored as JSON workflows on CERN Analysis Preservation and rerun using REANA. An analysis is wrapped inside a *Docker container*.

1.11 README

To build this documentation with Sphinx you have to follow a few simple steps.

1.11.1 Setup the Environment

If you only want to build the docs and you do not want to run the code or you want to keep these tasks in separate environments, continue with the following.

First, install the necessary requirements:

```
python python-virtualenvwrapper
```

Second, clone the repository (or your own fork) if you have not done so already:

```
git clone https://github.com/cernanalysispreservation/analysispreservation.cern.ch.
↪git cap
```

Third, create the virtual environment:

```
mkvirtualenv -p /usr/bin/python2.7 capdocs
```

and install Sphinx:

```
pip install Sphinx
```

Now you are all set. Whenever you want to build your docs in the future, just follow the below instructions.

1.11.2 Build the Docs

To build the docs, switch into the docs folder inside your repository folder

```
cd ~/PATH_TO_YOUR_CLONED_FOLDER/cap/docs
```

and run

```
make html
```

If that does not work and you do not see the `(capdocs)` in your terminal as such:

```
(capdocs) [USER@COMPUTER cap]$
```

then do the following:

```
workon capdocs
make html
```

1.11.3 Spell-Check the Docs

One possibility to spell-check the docs directly from your command line is to install:

```
hunspell hunspell-en
```

and run:

```
find . -type f -name '*.rst' -exec hunspell -d en_GB -l {} \;
```

from within the docs folder on your command line. This will give you a list of words possibly spelled incorrectly.

CHAPTER 2

License

Index