Template for PHENIX Analysis Notes

Gabor David[a], Yasuyuki Akiba[b]

[a] SBU / BNL
[b] RIKEN

Abstract

This write-up gives some guidelines/suggestions how to structure analysis notes, what information to include and to what depth. It is not meant to be a fill-out form or checklist, the more so, since no two analyses are similar. Instead, it's trying to convey the spirit in which notes should be written. Namely, always ask yourself: "Had I not been doing this analysis myself, would I be able to reproduce it and come to the exact same results based solely upon what's written in the analysis note?" If the answer is negative, please keep working on it. A positive answer means at the very least three things: 1/ all inputs (data, calibrations, recalibrators, deadmaps, ect.) are clearly defined (timestamp, version) or stored 2/ all codes, macros, scripts etc. used in the analysis are preserved safely, long-term (not in a volatile private directory) and 3/ the flow of the analysis is clearly described along with all auxiliary information needed to repeat it.

Contents

General guidelines

The overwhelming majority of PHENIX Analysis Notes are accompanying new physics results extracted from data. Before making them public, as preliminary or final, the submission of a comprehensive analysis note is a formal requirement, too. In reality, the amount and depth of information in the actual notes is quite disparate. While there's no way (and no point) to squeeze notes of the most diverse content into a single "checklist", or "fill-out form", we think some minimum should be adhered to. This template is attempt to guide you what this minimum should be (more information is always welcome!). Also, while writing your note, you should adhere to the following principles.

- **Reproducibility.** The purpose of an analysis note is to ensure that our published results are reproducible by **any** reasonably knowledgable PHENIX person. Reproducibility means that at the very least you give a full list of runs analyzed, production tag, analysis (train) code location and tag, analysis macros with location, QA criteria, cuts applied and procedures described comprehensibly. Err on the side of giving too much detail. **Omit** information referring to an earlier note **only** if that analysis did the **exact** same thing as you **and** if the description in it is truly comprehensive. Even a **non-expert in your topic** should in principle be able to **repeat** your analysis and come up with the **same result**.

- **Provide your own systematic error estimates.** Unless you are analyzing the exact same data with the exact same code, cuts and procedures for the exact same physics quantity (but then, why would you do that???) your systematic errors are **not** the same as your buddy's. If you estimated them independently and came up with the same number - that's fine, but show (prove) it! Statements like "X got this number so this is my error, too" are not acceptable.

- **Explain differences from previous analyses.** If a similar analysis has already been done before, and now you come up with different results or systematic errors, include a section in which you analyze/explain why did things change (hopefully improve...)

- **Point out problems.** Be honest. Saying that you didn't find the solution for a particular problem (and actually pointing out that problem instead of glossing over it) is both honest, ethical - and useful for the collaboration. Pretending that it doesn't exist, or that you solved it, even when you yourself know that's not true is the exact opposite. - If you found a mistake in an earlier analysis note, point it out, politely!

- **Start early.** Ideally analysis notes are written **in parallel with the actual analysis**, or at least when you are no more than halfway through. Yes, that means sometimes re-writing text, correcting numbers a dozen times, but at least at the end what you wrote down will be correct - which is not necessarily the case

when you just finished your analysis on time to meet a deadline but for approval you still have to submit an analysis note...

- **Make your note readable**. Give meaningful section (subsection...) titles and make a table of contents. Some otherwise excellent notes are lacking this "feature". Ever got frustrated trying to find two relevant sentences or a crucial number in 30-60-170+ pages of unstructured and un-indexed text? Also, since analysis notes are often long and sometimes hard to navigate, it is a good practice to give the exact location (pages) *within* a reference ("as described on pp. n-m of ANxxx"). You actually looked it up, right? Then this shouldn't be a burden.

- **Catch the spirit.** Analyses are quite complex, and it is hard, if not impossible, to write a set of formal rules, a checklist, which make sure that your analysis will indeed become fully reproducible. If in your case something more is needed than described here, please, please, add it! **Science is reproducibility**, and you are a scientist, aren't you? So please, act like one.

Selected "no-no"-s

To avoid being beaten up or lawsuits, here's the disclaimer: none of these examples are taken from actual notes, I just made them up. Unfortunately similar things can sometimes be found in actual notes.

- "We evaluated error X by a fast MC and it was negligible". *(Which fast MC? What parameters, input? How did you evaluate? How much is negligible?)*

- "We discarded runs with a bad RP distribution." *(Quantify what's "bad"!)*

- "We extracted the signal the same way as in Run-2". *(At the very least give a reference where the Run-2 analysis is described in detail. Second, it's hard to imagine that you did the exact same thing. I hope you didn't, that you learned in the meantime, your procedures became more sophisticated. But then they are different.)*

- "In this step we used NN's code/libraries". *(Using other people's libraries – unless, of course, they are recognized part of the official framework – is frowned upon. If absolutely unavoidable, specify the exact location and version, but a better solution is to copy the source and make it part of your own backup (HPSS or CVS). This also protects you if later NN decides to change something in his/her code.)*

- ... etc.

1 Analysis Organization

This is the section where you describe how your analysis is organized. It serves as a concise list, a catalogue of the steps of your analysis, the location of all necessary input and output files as well as all necessary codes, macros to do your analysis from scratch. It is meant to enable a third person to look later into your analysis and have access to everything needed to reproduce your results, if necessary. Basically it is about preserving the **data** and **software** necessary. Detailed description and explanation of the individual analysis steps – the **know-how** – comes in the later part of the analysis note where we still suggest some of the information to be repeated.

This section should include the following information.

A flowchart of the major analysis steps with code locations

The flowchart is a birds-eye view of your analysis flow, with references to the basic working directories. "Code location" here means your working directories on RCF. Ideally these all should branch off a single **base directory** (which also makes backing up the "snapshot" – see below – in HPSS easier). Important: if you did some of the work locally (e.g. on your laptop), you have to **migrate** the codes and relevant files **to RCF** when you are finished, and make sure it works there, too.

Examples of directories to be specified:

- Directory for Taxi code
- Directory to run Taxi
- Directory of the Taxi output
- Directory for Simulation code
- Directory for Simulation output
- Directory for analysis code and macros that analyze the Taxi output
- Directory for analysis code and macros that analyze the Simulation output
- Any other directory used during the analysis
- Directory for the final data file(s) and macros to produce physics plots.

We suggest you write these directories as they are at the time you finished your analysis, and you reference them in the later part of the analysis note. After they are backed up in HPSS, part or all of it can be removed.

Location in HPSS of your analysis snapshot

We request that a snapshot of your analysis, namely, all of the directories and the files in the directories described above, are backed up in HPSS.

Before the backup, make a good faith effort to **clean up** the directories (such that someone later looking at it isn't unnecessarily distracted). Log files, for instance, can be deleted. However, make sure that they contain **all those files** that are needed to repeat the analysis from scratch. **Avoid symlinks!** If you

aggregated the taxi output (e.g. with *haddPhenix*), you can make a text file of the run numbers processed (runlist), then delete the individual run outputs. Same for simulations. However, keep a small fraction of the original files (and include them in the backup). Then please document the location of your analysis snapshot in HPSS here.

Note: At some point in the future the Data Preservation Task Force is planning to centralize the location of those analysis snapshots into one place. In order to make this easier, in addition to documenting the HPSS location of your snapshot in the analysis note, please also send an email indicating this location to Chris Pinkenburg, Maxim Potekhin, Takahito Todoroki and Gabor David.

Dependencies. If you are aware that your code uses libraries other than the ones created from your own code or are part of the core PHENIX software, please point out this fact here. Ideally the source code for those other libraries should also be included in your backup.

Remember, working directories are handy, easy to access, but they can easily come and go, be deleted, altered, so a copy of the snapshot of your analysis - files, codes, macros - at the time of preliminary request or the final publication should exist in HPSS/CVS.

If you don't know how to back up a directory tree to HPSS, you'll find instructions at the end of this section.

Brief summary of your analysis

State the physics observable analyzed, the basic method and the subdetectors involved. Then please write a brief description (overview) of the steps of your analysis and in which your working directory each step is done. These are essentially bullets explaining the flowchart and adding some basic information, with pointers to working directories. Detailed description comes in later sections of the analysis notes.

Data Set

- Run, beam species, energy, trigger(s)
- Good run list and its location. (Don't put the good run list itself here. If you want to include it in the analysis note, too, put it in an Appendix)
- Number of events analyzed and the corresponding number of MB events or integrated luminosity

Taxi

- Taxi code and its CVS location, including version number
- List of recalibrators used for the analysis. Are all recalibrators public? If there is any private recalibrator used, please list them and their location.

- List of files like efficieny map or data for recalibrator
- Taxi output

Simulation

Simulation code and any related files

- List of simulation code used for the analysis
- Event generator(s), including tuning/parameter files
- PISA setup file
- Tuning of PISA
- Output of simulation (PISA, PYTHIA, etc)
- List and location of code to analyze the simulation data and its output
- Any other pertinent information

List of major intermediate analysis files

There are several steps in the analysis from Taxi run to the final data file. You will produce intermediate analysis file in each step. Please list major intermediate files produced by the analysis. Remember: reproducibility means that **each** step of your analysis process can be unambiguously traced. For example

- Code to produce the raw data histogram of pi0 from Taxi output
- ROOT file of the raw data histogram of pi0 pT distribution
- ROOT file of PISA simulation for pi0 efficiency calculation
- Embedding code/macro
- Code to analyze the PISA simulation file to produce pi0 efficiency
- ROOT file of the efficiency vs pT of pi0
- Macro to calculate the cross section from the raw histogram and efficiency
- Any other pertinent information

Final data file and plotting macro

We now request that final physics data for figures for a journal paper and preliminary plot should be put in ROOT files, and the figures/plots are produced by plotting macros from these final data files. We request that all of the ROOT files of plotted data and the plotting macros are in a single directory. Please list the final data files, plotting macros, and their location here. Please indicate which macro will produce with plot from which data file.

In addition you have to provide your **final data points** with uncertainties in **text files** (ASCII), and specify their location. These files are needed because 1/ we have to post the data on our public website 2/ we want to upload them in HEP databases. (While strictly speaking this is not a requirement for preliminary data, it is good practice to provide ASCII files from those, too.)

Example:
The data file and plotting macros of the preliminary request is located at

`/phenix/u/analyzer/analysis/prelim2019.10`

`DataFile.root` The data file of all plots requested for preliminary
`plot_xs_pi0.C` Macro to produce Plot1 (cross section of pi0).
`plot_RAA_pi0.C` Macro to produce Plot 2 (RAA of pi0)

The ROOT files of the plotted data and the plotting macros should be backed up in HPSS as a part of the snapshot of your analysis.

We request that the directory of the ROOT files of the plotted data and the plotting macros should be under your user directory (`/phenix/u/analyzer`) so that it is in RAID part of RCF filesystem. We request that this final data directory should be "frozen" at the time of preliminary request or final paper publication, except for editing of the plotting macros for cosmetic changes of the plots. Important: this directory should be duplicated on the rcas disk, and be part of your analysis snapshot, too (backed up in HPSS).

How to put a directory tree in HPSS?

First make a .tar file of the directory tree, preferably on the PHENIX scratch area (if you don't already have a directory there, you can create one)

```
cd /gpfs/mnt/gpfs02/phenix/scratch/david
```

then do the equivalent of this

```
tar -cvM -f plhf1_david_taxi.tar /gpfs/mnt/gpfs02/phenix/plhf/plhf1/david/taxi
```

While sitting in this (scratch) directory, log in to HPSS (hsi command, the password is the one you use for rcf3, rcas)

```
[david@rcas2064 david]$ hsi
Warning: "Network Options" section is empty in HPSS.conf file
nd_krb_preexist_auth: No credentials cache found retrieving principal
name from cred.cache
nd_kerberos_auth: No credentials cache found on krb5_mk_req call
Kerberos Principal: david
Password for david@SDCC.BNL.GOV:
Username: david  UID: 1113  Acct: 1113(1113) Copies: 1 Firewall: off
[hsi.5.0.2.p3 Mon Jun 29 16:29:32 EDT 2015]
? pwd
pwd0: /home/david
```

You end up in your HPSS directory. Create a subdirectory for the new full backup (in this case it is my old taxi directory), then "mput" the tar file in HPSS (per default "mput" will copy files from the directory where you have been when you logged in to HPSS with the hsi command):

```
? mkdir taxi_20190828
mkdir: /home/david/taxi_20190828
? cd taxi_20190828
? mput *
mput 'plhf1_david_taxi.tar'? ([Y]es,[N]o,[A]ll,[Q]uit) y
Active:   1, Queued     0 [In-flight: 193.25GB Xferred: 1.10GB 0.57%] [Done: 0 F
```

and it will keep showing you the progress how the file is uploaded in HPSS. Since the file first goes to a huge buffer-disk, the transfer is pretty fast. It ends with this message, when you log out:

```
mput  'plhf1_david_taxi.tar' :
'/home/david/taxi_20190828/plhf1_david_taxi.tar'
( 193246924800 bytes, 220612.0 KBS (cos=13))
? ls
/home/david/taxi_20190828:
plhf1_david_taxi.tar
? q
[david@rcas2064 david]$
```

That's all, you are back on rcas, in the scratch directory. Be nice and delete the .tar file right away (don't wait for the automatic deletion).

2 Introduction

Define the purpose and scope of your analysis. Describe (briefly!) the physics you are after and how/why is the signal you extract relevant to it. If this is not the first time such signal is analyzed, refer (briefly!) to earlier PHENIX work and point out what's different (method? dataset? statistics? better simulations? new insight in some bias? bugs found and fixed? ...) Describe (briefly!) the improvements you made and the major unresolved issues.

In what follows, keep in mind that your analysis has to be **reproducible**. *If you fulfilled requirements described in Sec. 1 it means that the* **software** *environment is preserved. But that's obviously not sufficient. Now you have to make sure that the* **know-how** *is preserved too, the "recipees" how to do the individual steps, proper sequence, dependencies are spelled out. In some cases this is trivial, but it can also get very tricky, interconnected, iterative; you have to make sure, that a third person can understand, what to do with the stored software.*

3 Data set, QA, general cuts

Decribe precisely your input dataset (a table of all runs analyzed should appear in the Appendix [1]*): data taking period, master DST/pDST production tag, train number. Describe your run selection (QA) criteria, preferably with justification. Plot the relevant global quantities (centrality? reaction plane? multiplicity? ...) from the runs passing your QA.*

If you have event selection criteria, list and justify them. Show the relevant global quantities before and after selection. If you use triggered data, show the turn-on curve.

Summarizing in a table what fraction of runs/events were eliminated by different cuts is good practice.

Plot live/dead areas (maps) in the relevant detectors. Indicate if they influence acceptance in a p_T- or charge-dependent way. Indicate if they are fixed over the dataset or run-dependent.

4 Data Analysis

4.1 Analysis chain

List the steps of your analysis (typically this will be a much longer list than the one in Sec. 1) and point to the subsections where they are described in detail. If you are doing something new/unusual, point out this fact here.

As said before, there's no way to write up a general, mechanical "checklist". Use your judgement and always try to think along the lines: "If I were the outsider, would I fully understand what has been done?" If not, keep working on the analysis note.

[1] Yes, this still leaves the question of segments open, but that would be overkill

4.1.1 Analysis step 1

Give a meaningful title to the subsection. Provide a reference to the directory where this step has been done and what macros, scripts were run, with what parameters. Define your cuts precisely, and plot its effects (before/after). For log distributions consider ratios. If you make a fit, always give the precise functional form, its parameters and plot the data/fit ratio.

4.1.2 Analysis step 2, ... n

In general, plot intermediate results. It is much easier to believe not only your ideas, but the actual work you have done if one sees how your raw data evolved in the individual steps into the final, fully corrected results.

4.1.3 Simulations

Your simulations should also be reproducible: provide all pertinent information (generator, acceptance cuts, ranges, ...). If simulation output has to be modified (like an additional smearing), explain. Provide plots that show how well simulations describe real data (like simulated and real peaks, widths). Provide location and tag of the embedding code (if any).

4.2 Systematic errors

Make your own estimate *for each error related to your own analysis, except for "external" quantities you take from someone else (centrality, reference spectrum...) - for those give exact references. If you claim some error cancels in a composite quantity computed from your and an "external" result (like a p+p denominator in a heavy ion R_{AA}) make sure and prove with a reference that the relevant parts of the analyses were done the exact same way (often they were not!).*

4.2.1 Error source 1

Describe the error source, explain, what type it is (A,B,C,...) and whether it is centrality-dependent, absolute or relative. Explain how you estimated its contribution, and whether your estimate is full extent, σ or something else. Include a plot from which the reader can pass his/her own judgement on your estimate. Always give a specific number or upper limit (not just "it's small"). Your errors should be clearly defined here (if necessary, listed in a table) for all data points.

Confessing that a particular estimate is "soft" is not a shame (even if the PWG grilles you for it).

4.2.2 Error source 2, ...n

If possible, for p_T-correlated errors provide an estimate how is it (anti)correlated; a function is even better. Justify it.

4.2.3 Summary of systematic errors

Provide a table that summarizes your systematic errors. If such table would become too big, you may want to select only a few points, representative of your entire range (p_T, centrality,...).

4.2.4 (If relevant: changes in error estimates since earlier results)

If this is not the first analysis of its kind (preliminary to final, re-analysis of a dataset, etc.) and if your systematic errors differ significantly from the ones found before, discuss in detail what changed (method? assumptions? new insight? etc.) and why do you think the new estimate is better than the previous one.

4.2.5 (If relevant: how to improve systematic errors in the future?)

This could be one of the most useful parts of your note. During your analysis you often got frustrated, because you couldn't improve on something due to lack of specific data, software capabilities, unresolved detector problems, etc. Discuss them, particularly if they turned out to be your dominant errors, and suggest improvements. Try to be specific, beyond generalities like "we should do more simulations" or "we need better hadron rejection".

5 Results

Plot all your results with clear captions. If there are too many (say, more than 10-12 pages) break them up into subsections for easier reading/navigation. If you use "external" data (like reference spectra), plot them separately, too. If previous results of the same quantity exist, plot old/new, and if there are big discrepancies, describe briefly the reason (even if you discussed it already in detail in the analysis section).

5.1 Discussion, comparison to theory

If possible, relevant and available, compare your results to theories, plot and discuss the results. This is not a paper to be published, so feel free to wander off in less-than-mainstream directions as well. They might actually give someone an idea...

6 Appendix

6.1 Data tables

No results without data tables! List errors separately (statistical, type A, B, C) and the total. If you have many tables, breaking them up into subsections is strongly recommended. Any analysis note without data tables will be immediately rejected.

6.2 (Runlist)

References

[1] Your reference http://www.phenix.bnl.gov/yourreference/ref1.pdf