FAIR Data for Complex HEP Datasets

FAIROS-HEP Kick-Off Workshop - 8-10 Feb 2023

Kati Lassila-Perini
Helsinki Institute of Physics - Finland
CMS Data preservation and open access coordinator

1

CMS Open data – FAIR?

Findable – Accessible – Interoperable – Reusable

FAIR? My interpretation… (see also Tibor's talk)

FINDABLE

Do you know where to look for them?

Can you find what you need?

ACCESSIBLE

Can you download them?

Are they in some common format?

Do you have the tools to open the data files?

INTEROPERABLE

Do you know how to use?

Can you make new research with them?

REUSABLE

FAIR? My interpretation...

FINDABLE

Do you know where to look for them?

Yes, CERN open data portal

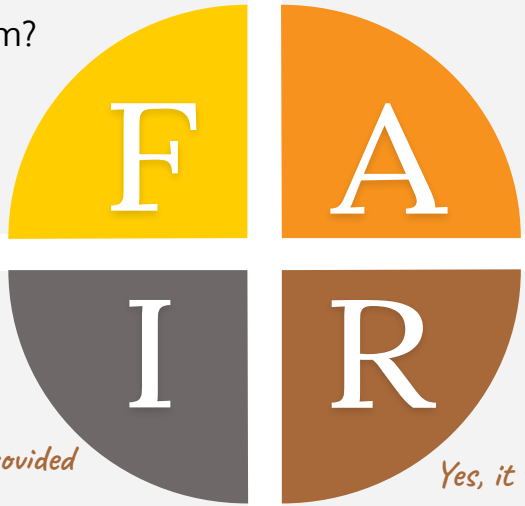
Can you find what you need?

Yes, there are search functions

ACCESSIBLE

Can you download them?

Yes, or they can be streamed with XRootD



Are they in some common format?

No, not really...

Do you have the tools to open the data files? *No, but they are provided*

INTEROPERABLE

Do you know how to use?

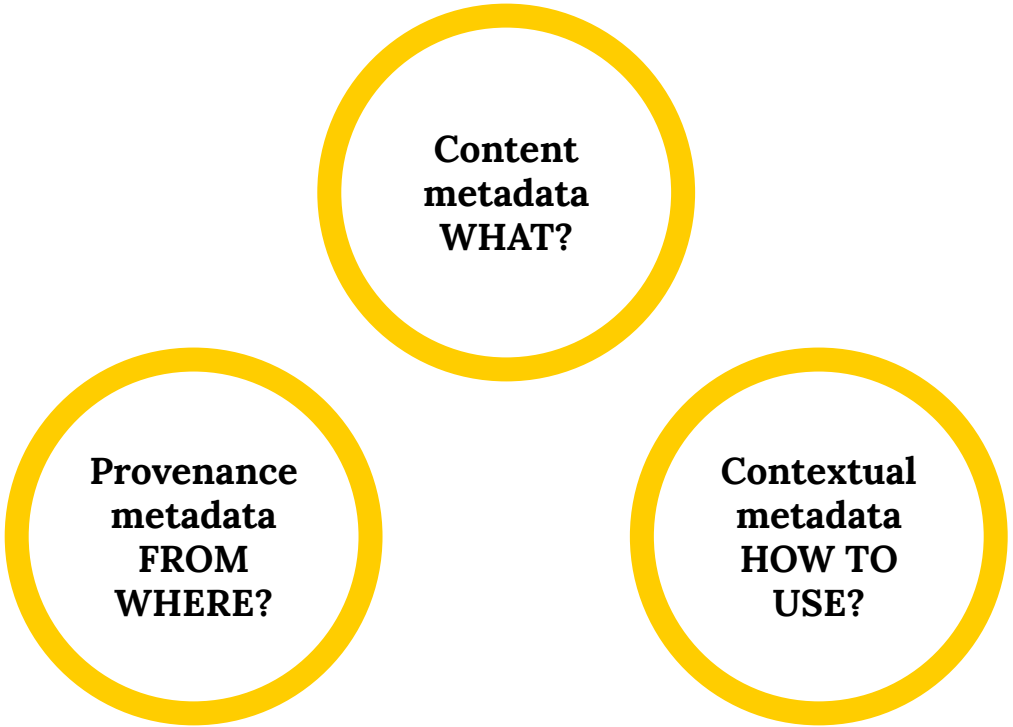
There are instructions to get started...

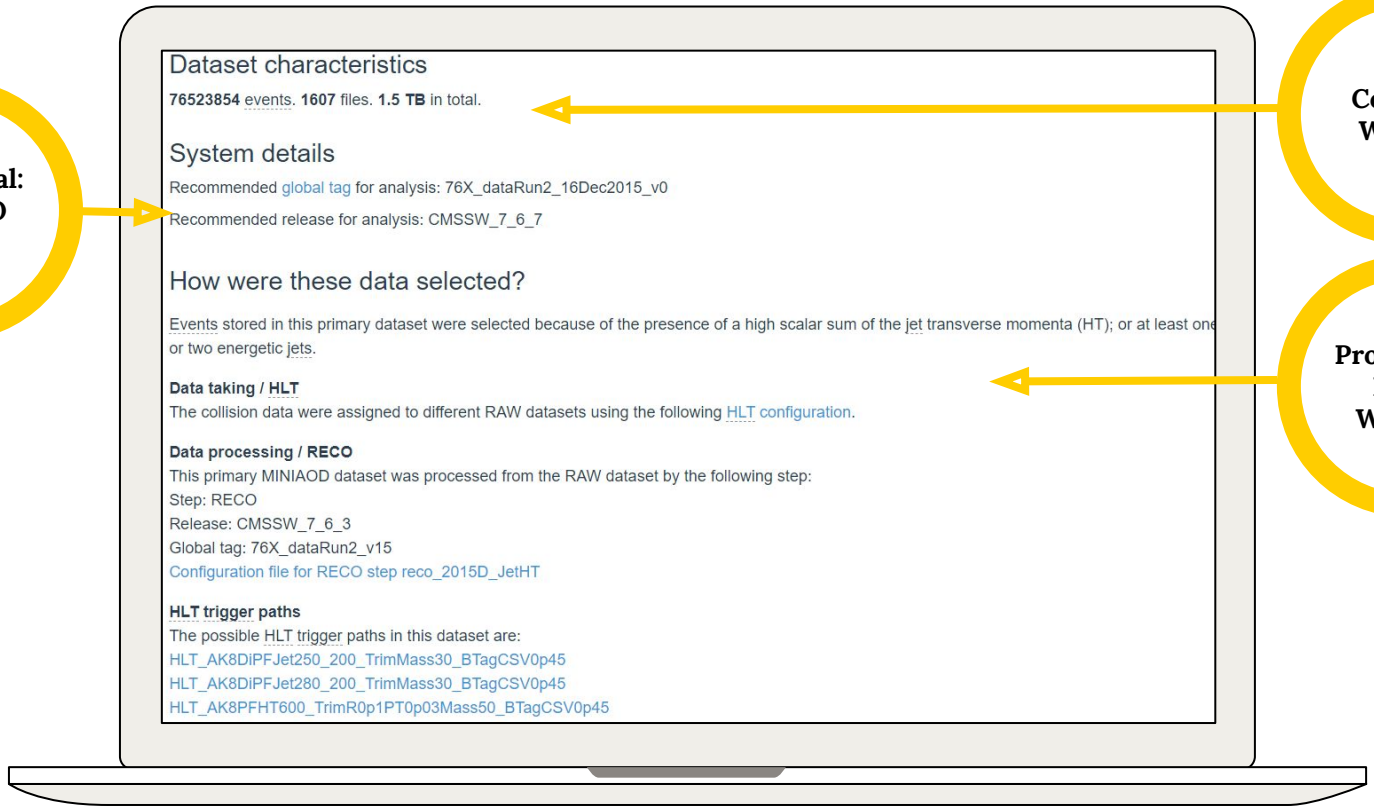
Can you make new research with them?

Yes, it takes at least as much as for the CMS people

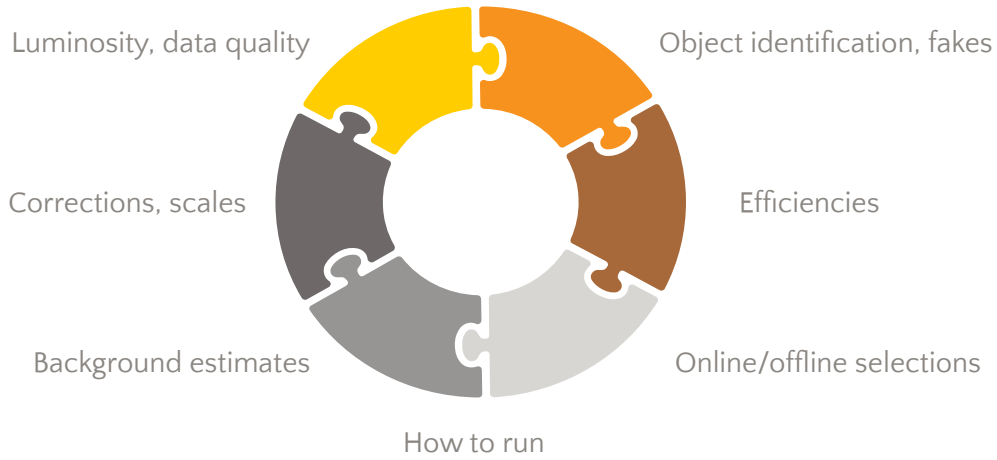
REUSABLE

Metadata for complex research data

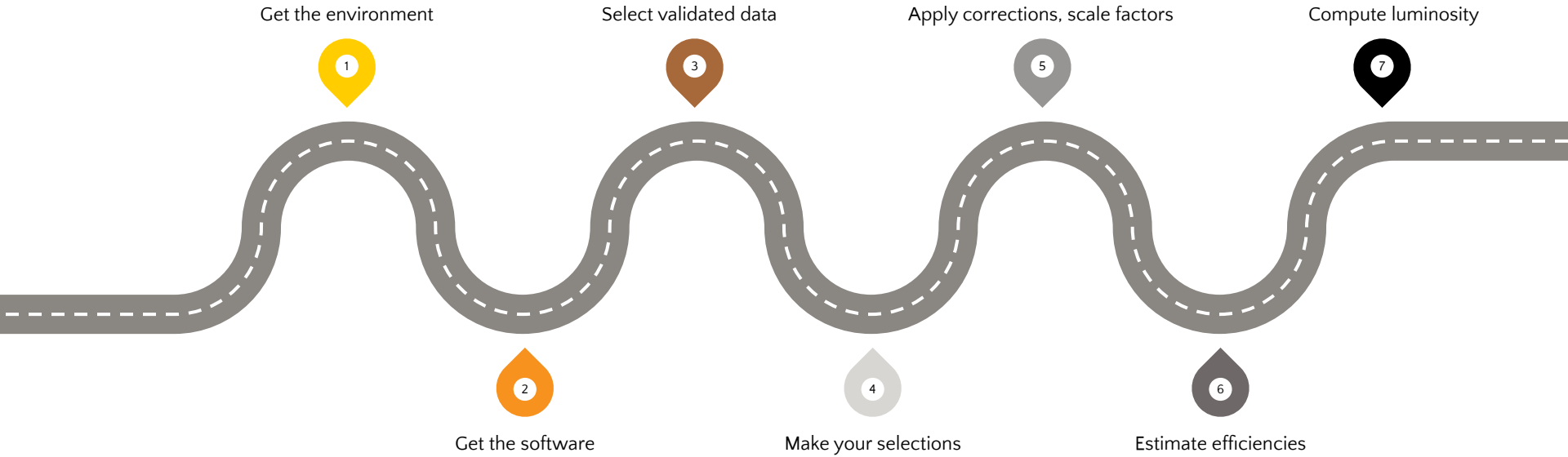



Dataset characteristics
76523854 events. 1607 files. 1.5 TB in total.
System details
Recommended global tag for analysis: 76X_dataRun2_16Dec2015_v0
Recommended release for analysis: CMSSW_7_6_7
How were these data selected?
Events stored in this primary dataset were selected because of the presence of a high scalar sum of the jet transverse momenta (HT); or at least one
or two energetic jets.
Data taking / HLT
The collision data were assigned to different RAW datasets using the following HLT configuration.
Data processing / RECO
This primary MINIAOD dataset was processed from the RAW dataset by the following step:
Step: RECO
Release: CMSSW_7_6_3
Global tag: 76X_dataRun2_v15
Configuration file for RECO step reco_2015D_JetHT
HLT trigger paths
The possible HLT trigger paths in this dataset are:
HLT_AK8DiPFJet250_200_TrimMass30_BTagCSV0p45
HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV0p45
HLT_AK8PFHT600_TrimR0p1PT0p03Mass50_BTagCSV0p45
Contextual: HOW TO USE?
Content: WHAT?
Provenance: FROM WHERE?

Contextual metadata - should cover a lot!

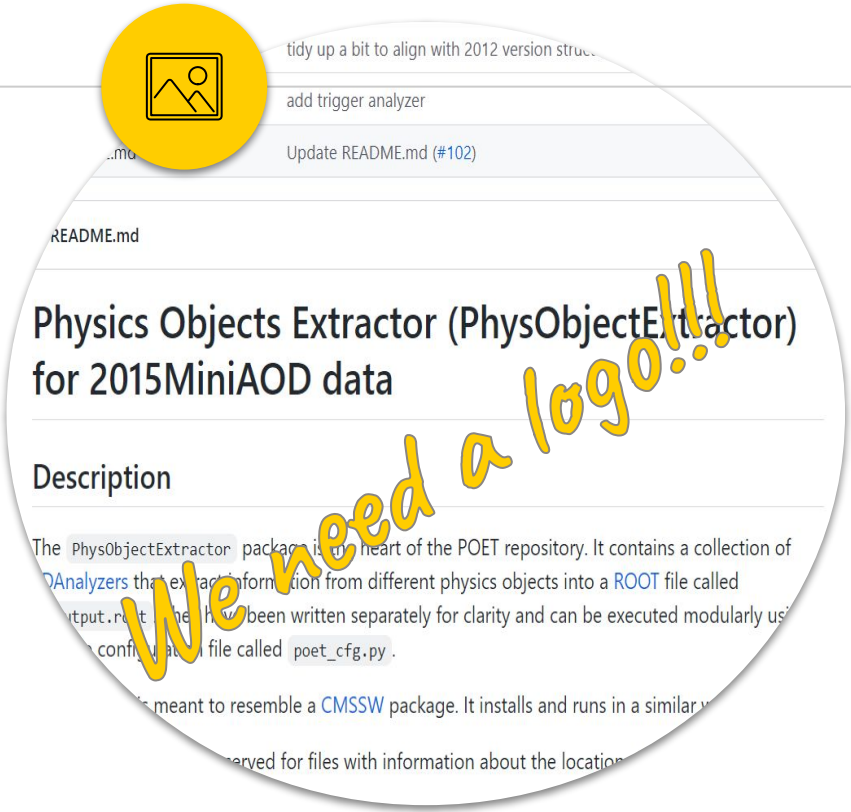


How to put this together?



Providing contextual metadata in a useful way

- Best through example workflows
  - automated, machine and human-readable
- Not easy to define (even with >1100 papers...)
  - partly because
    - analysis processes are complex
    - CMS data format supports a wide range of usecases
  - but also because we, as a community, have undervalued:
    - documentation
    - common tools
    - analysis code reuse.



THE example code: POET

Physics Objects Extractor Tool

Put together by many people in the CMS open data group from various sources within CMS.

See Julie's talk on Wednesday.

Not a negligible effort!

2

Open science - how to?

Example workflows do not come out of the blue - it has been a major effort to put them together for CMS open data.
A little interlude to what it takes to open science to happen

2.1 Roles & responsibilities

Open science - what it takes to make it happen

Knowledge preservation
Data and software skills
Time

Infrastructure
Resources
Guidelines

Policies
Agreements
Practices

2.2

Best practices

To preserve the knowledge at the time of active analysis

Efforts are needed

Package your analysis environment in software containers

Use continuous, automated testing

Capture the steps in your analysis

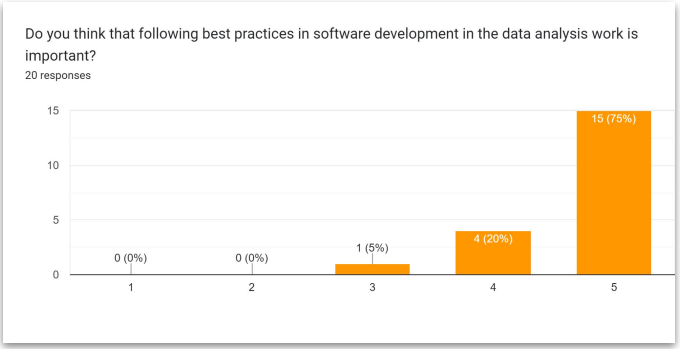
Define easily reusable workflows

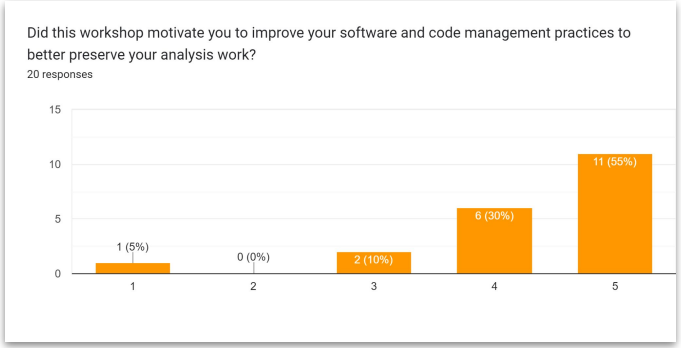
Manage your code in versioned code repositories

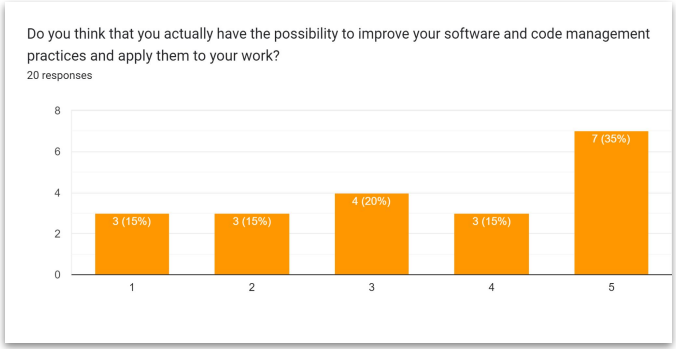
Document everything from the start

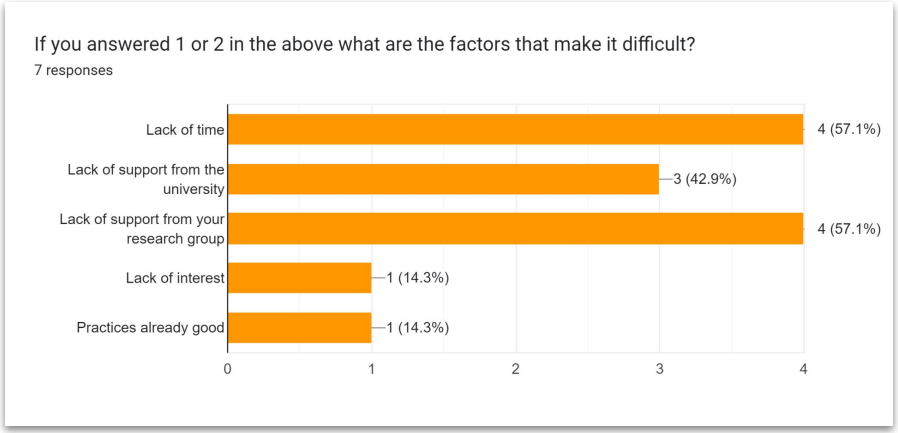
Best practices require time but they will pay off:

for the individual, for the group, and eventually, for open science!









Feedback from a recent Open data -workshop for PhD students in physics (outside HEP)

All parts count

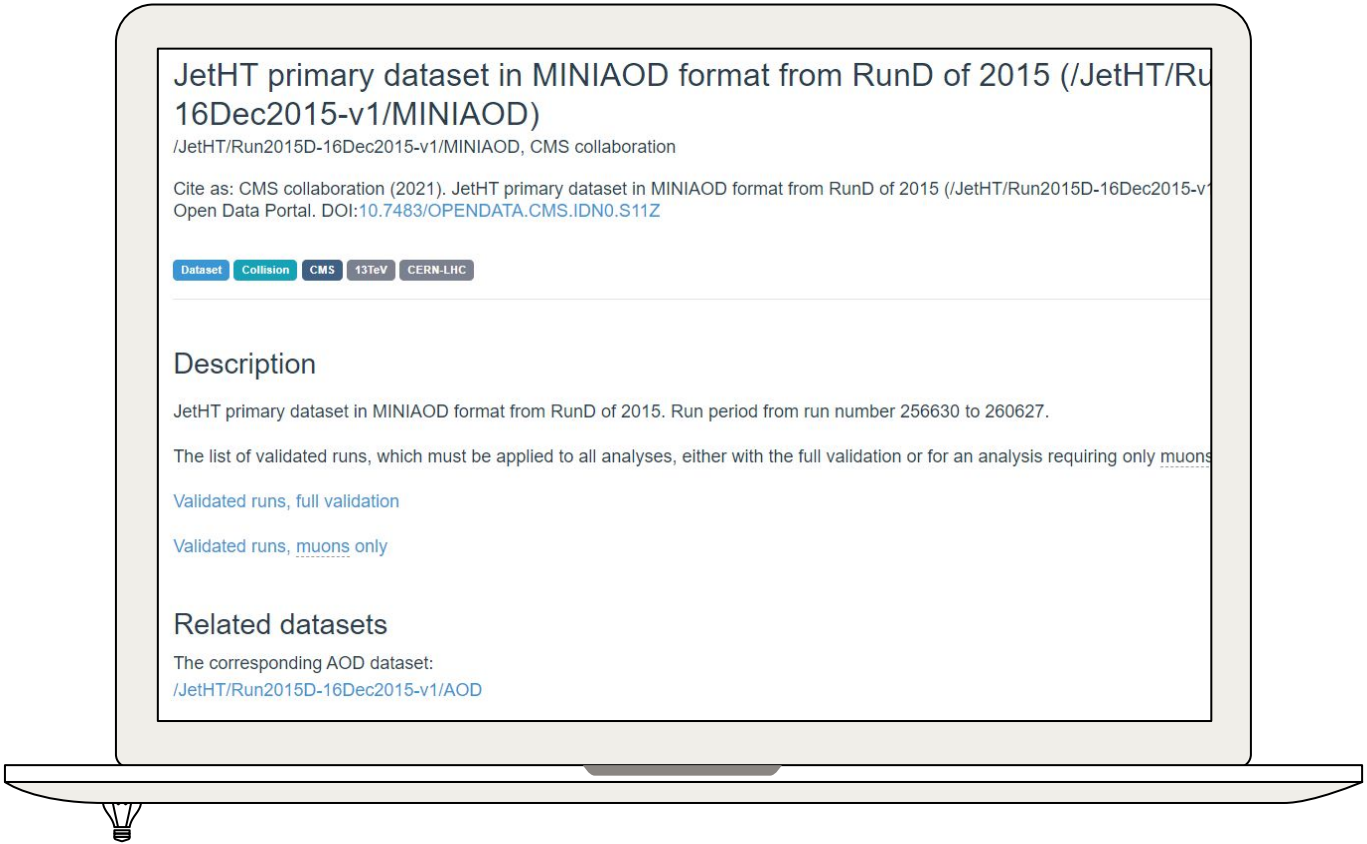
It is the researcher that has the key role in preserving the knowledge and needs support in doing so!

3

Assessing FAIRness with workflows

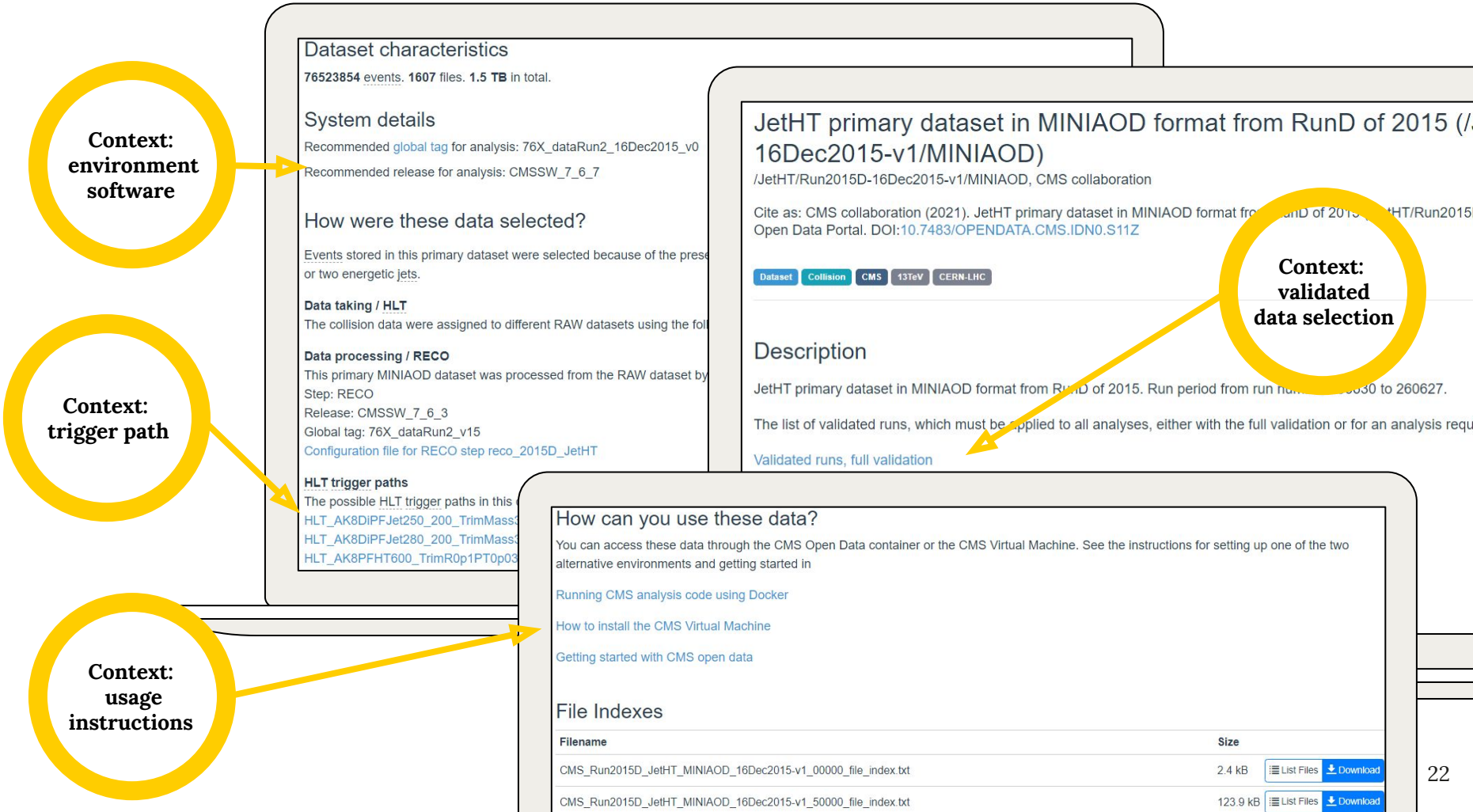
Automated workflows that concretely access, use and test the CMS open data metadata in the research context.

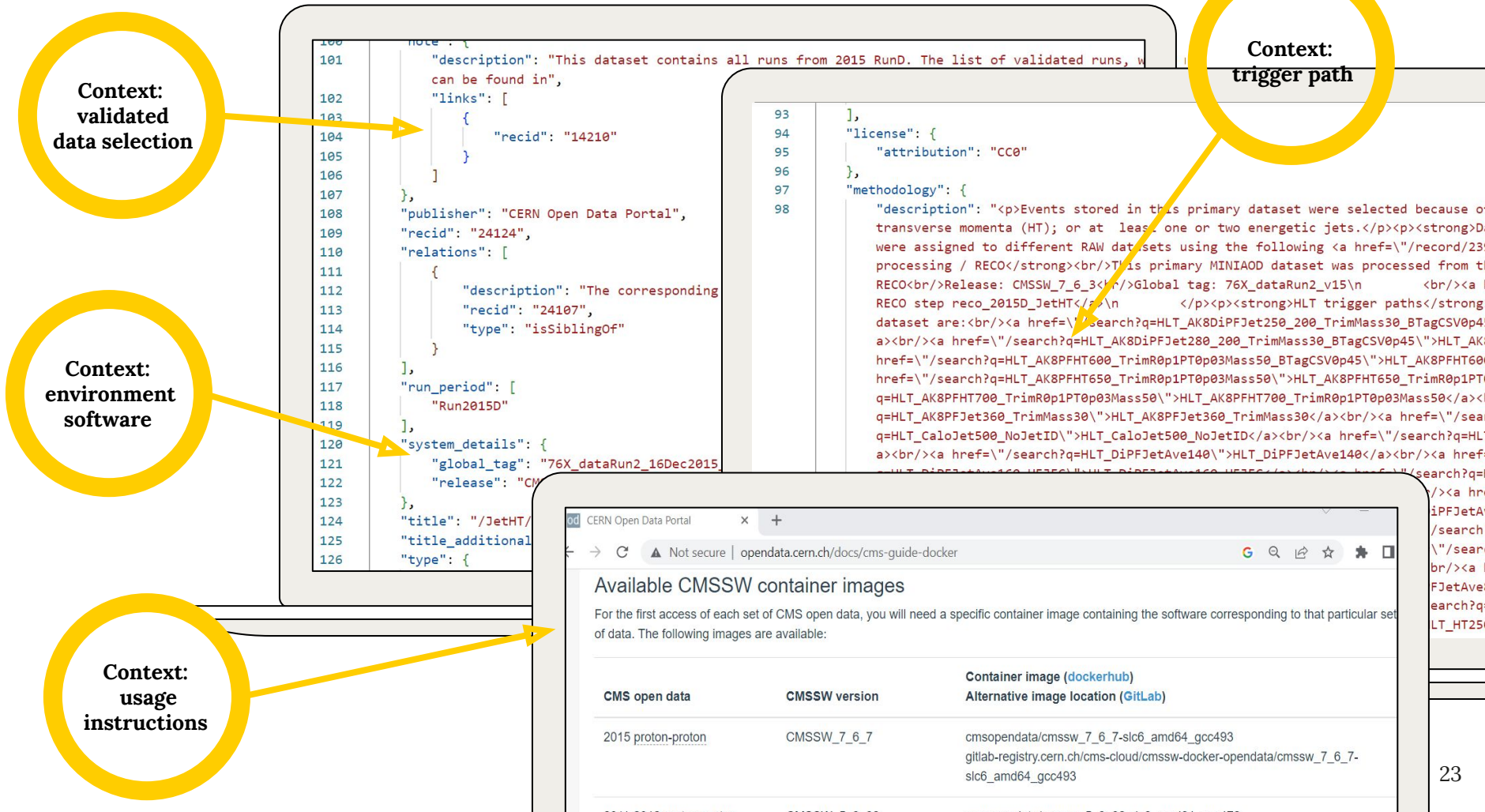

CMS Open Data workshop 2022!
The code for the workflow?
Assembled for the CMS OD workshops!
Watch on YouTube



Open dataset record display

Underlying metadata


Context: environment software
Context: trigger path
Context: usage instructions
Context: validated data selection
Dataset characteristics
76523854 events. 1607 files. 1.5 TB in total.
System details
Recommended global tag for analysis: 76X_dataRun2_16Dec2015_v0
Recommended release for analysis: CMSSW_7_6_7
How were these data selected?
Events stored in this primary dataset were selected because of the prese
or two energetic jets.
Data taking / HLT
The collision data were assigned to different RAW datasets using the foll
Data processing / RECO
This primary MINIAOD dataset was processed from the RAW dataset by
Step: RECO
Release: CMSSW_7_6_3
Global tag: 76X_dataRun2_v15
Configuration file for RECO step reco_2015D_JetHT
HLT trigger paths
The possible HLT trigger paths in this
HLT_AK8DiPFJet250_200_TrimMass3
HLT_AK8DiPFJet280_200_TrimMass3
HLT_AK8PFHT600_TrimR0p1PT0p03
JetHT primary dataset in MINIAOD format from RunD of 2015 (/
16Dec2015-v1/MINIAOD)
/JetHT/Run2015D-16Dec2015-v1/MINIAOD, CMS collaboration
Cite as: CMS collaboration (2021). JetHT primary dataset in MINIAOD format fro
Open Data Portal. DOI:10.7483/OPENDATA.CMS.IDN0.S11Z
Dataset
Collision
CMS
13TeV
CERN-LHC
Description
JetHT primary dataset in MINIAOD format from R
The list of validated runs, which must be
Validated runs, full validation
How can you use these data?
You can access these data through the CMS Open Data container or the CMS Virtual Machine. See the instructions for setting up one of the two alternative environments and getting started in
Running CMS analysis code using Docker
How to install the CMS Virtual Machine
Getting started with CMS open data
File Indexes
Filename
Size
CMS_Run2015D_JetHT_MINIAOD_16Dec2015-v1_00000_file_index.txt
2.4 kB
List Files
Download
CMS_Run2015D_JetHT_MINIAOD_16Dec2015-v1_50000_file_index.txt
123.9 kB
List Files
Download


Context: validated data selection
Context: environment software
Context: usage instructions
Context: trigger path
"description": "This dataset contains all runs from 2015 RunD. The list of validated runs, w
can be found in",
"links": [
"recid": "14210"
"publisher": "CERN Open Data Portal",
"recid": "24124",
"relations": [
"description": "The corresponding
"recid": "24107",
"type": "isSiblingOf"
"run_period": [
"Run2015D"
"system_details": {
"global_tag": "76X_dataRun2_16Dec2015
"title_additional
"type": {
"license": {
"attribution": "CC0"
"methodology": {
CERN Open Data Portal
Not secure | opendata.cern.ch/docs/cms-guide-docker
Available CMSSW container images
For the first access of each set of CMS open data, you will need a specific container image containing the software corresponding to that particular set
of data. The following images are available:
CMS open data
CMSSW version
Container image (dockerhub)
Alternative image location (GitLab)
2015 proton-proton
CMSSW_7_6_7
cmsopendata/cmssw_7_6_7-slc6_amd64_gcc493
gitlab-registry.cern.ch/cms-cloud/cmssw-docker-opendata/cmssw_7_6_7-
slc6_amd64_gcc493

First thoughts: to-do list

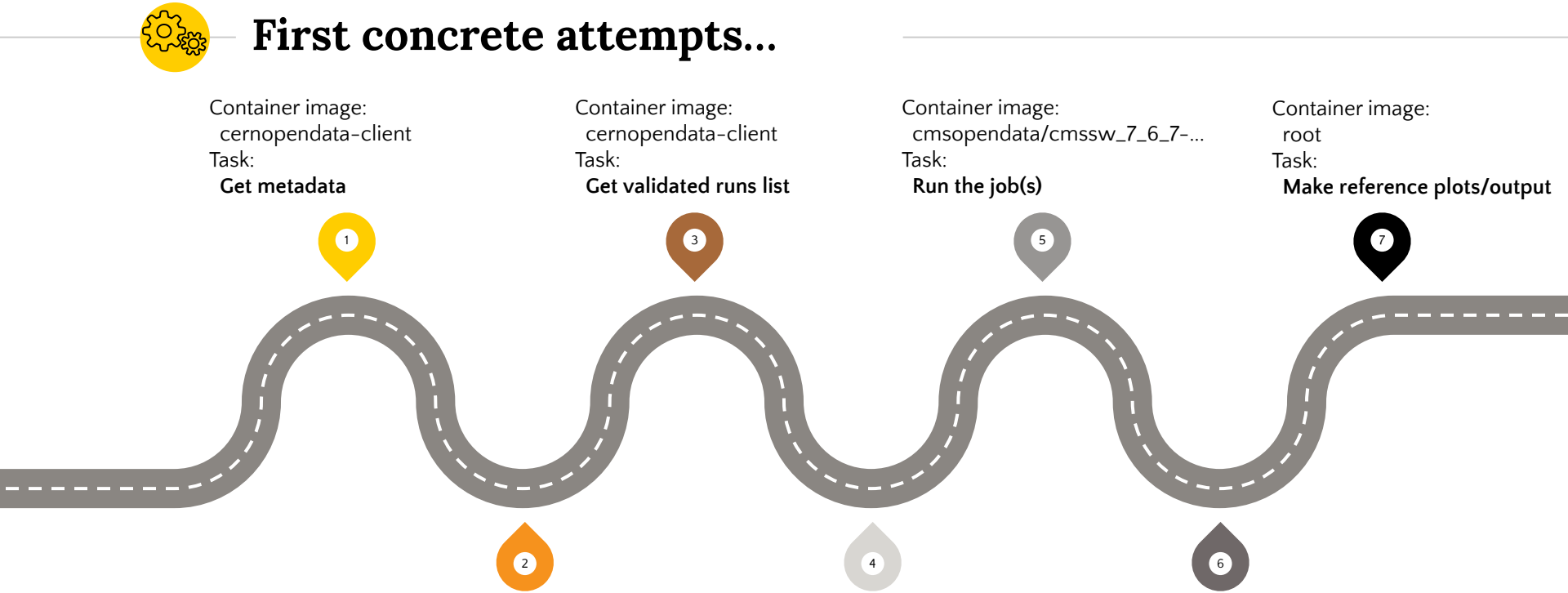
- The CODP metadata is display-oriented and some fields of interests contain HTML
  - Not easily machine-readable

⇒ Modify the existing CODP metadata structure

- Some contextual metadata is not exposed but embedded in the software or in condition data
  - Trigger prescales, corrections, scales etc

⇒ Will be addressed with nanoAOD...

First concrete attempts...

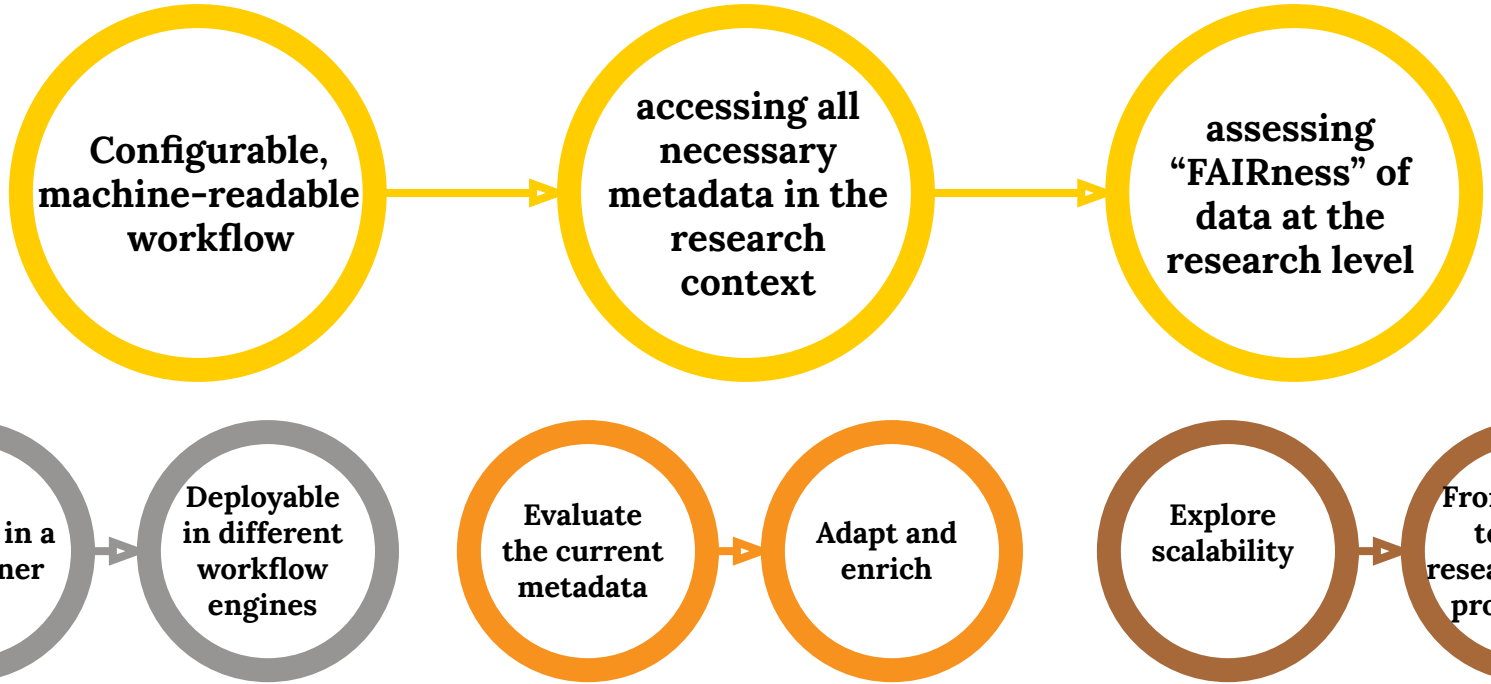


4

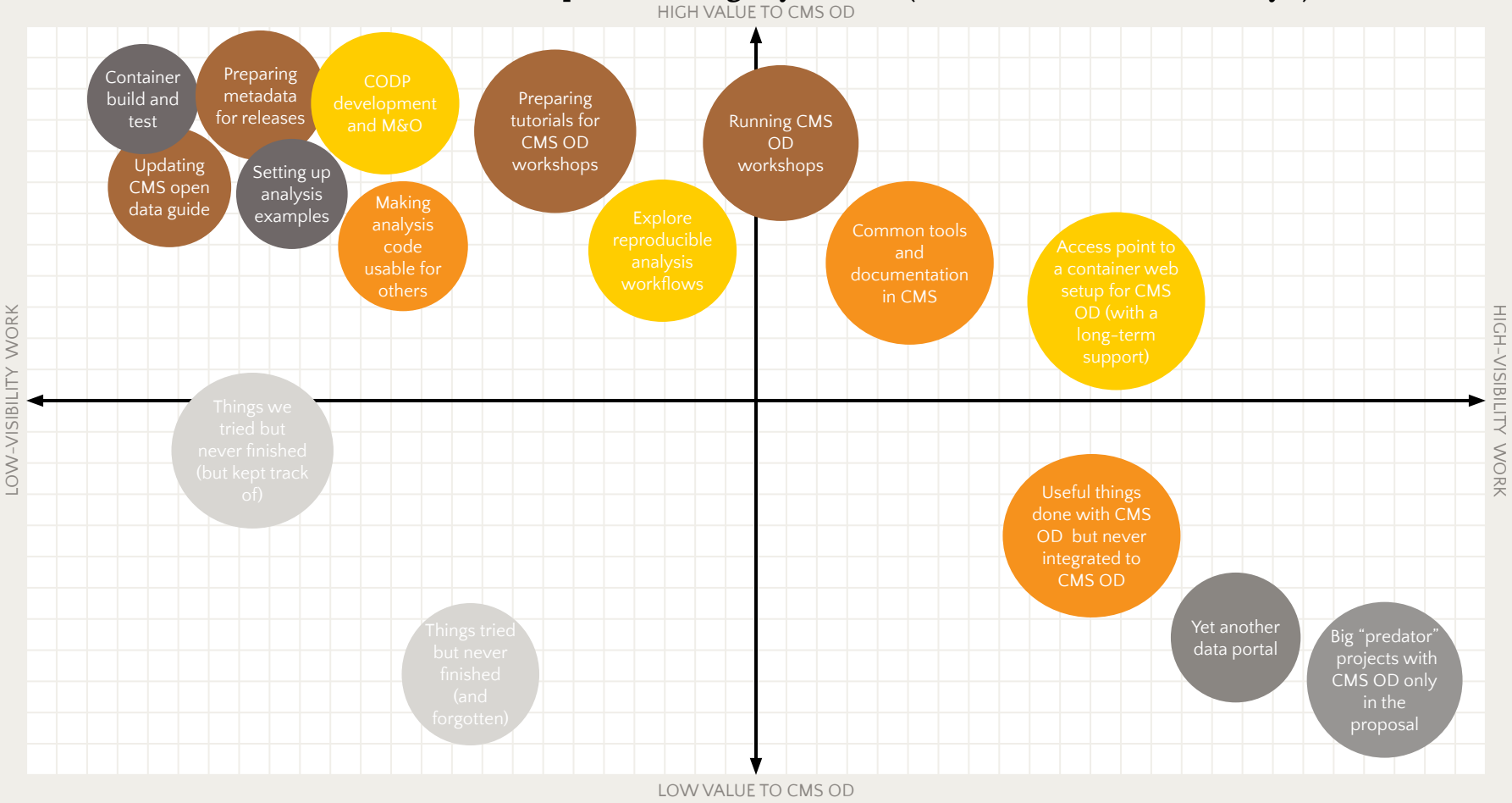
Conclusions

and some personal considerations...

Proposal:
CMS open data as a testbed



Personal considerations: Value to CMS open data vs “glory” Matrix (not to be taken too seriously…)



Thank you!

Any questions ?

And thanks to SlidesCarnival for this free presentation template