*Workshop on Open Science practices in Nuclear Physics, 5-6 December 2022*

# Path to Open Science @ AGATA

O.Stézowski,
as member of the AGATA Collaboration

Many thanks to the AGATA Data Processing Team

G.Baulieu, N.Dosme, J.Dudouet, S. Elloumi, Ph. Gauron, A. Goasduff, M.Gulmini,
A. Korichi,
J. Jacob, V. Lafage, E. Legay, P. Lejeannic, J. Ljungvall, G.Philippon, R.Molina, M. Roetto, M. Tauriga-Quere, N.Toniolo

# The Data Processing « Phase 2 » Time Line

International Review
KDP2 IN2P3 } ➡ New MoU AGATA Phase 2 (signed now)
➡ Data Management Plan required

Electronic Phase 0

Electronic Phase 1 and Phase 2
PRISMA, Trace, coupling to xDAQ …

Electronic Phase 2 only
Ancillaries …

GGP only

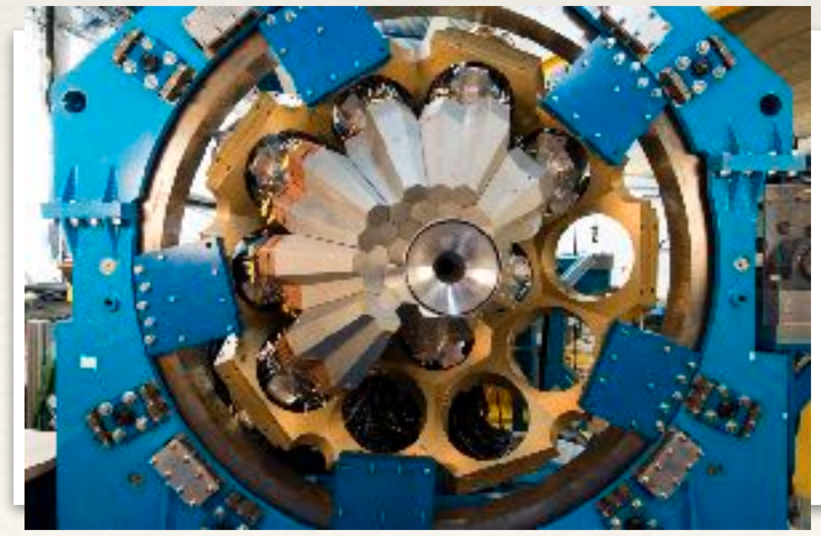
GGP only + V2

2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030

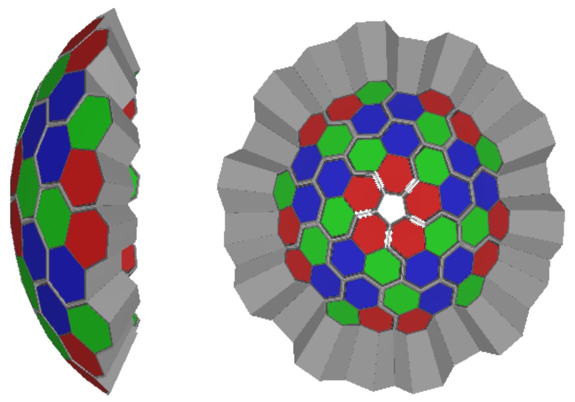
Dismounting
@ GANIL

Re-start
@ Legnaro

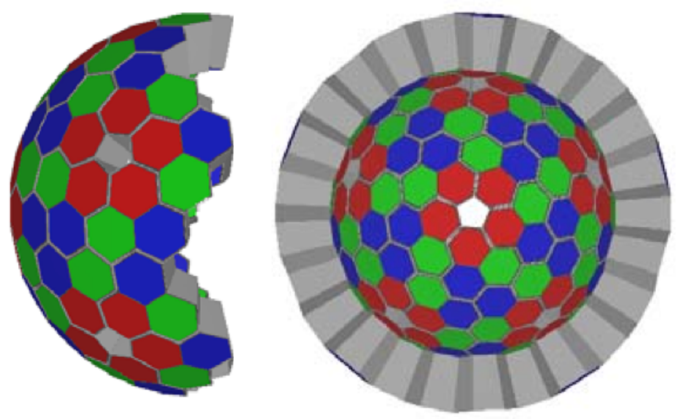
New review …



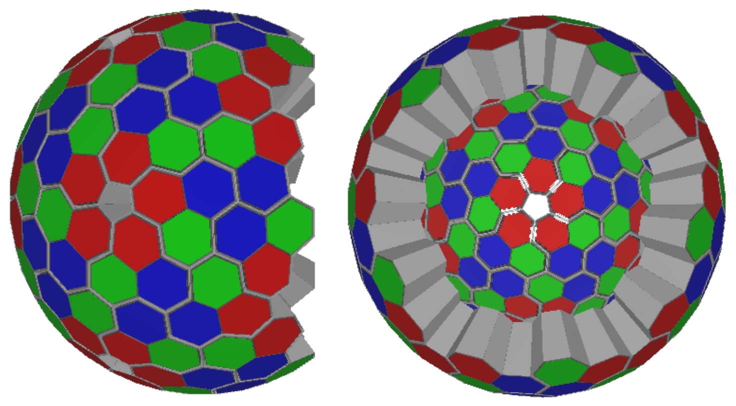
4PI/3
60 Ge
20 TC



2PI
90 Ge
30 TC



3PI
135 Ge
45 TC



# The Data Processing « Phase 2 » Time Line

International Review
KDP2 IN2P3 } ➡ New MoU AGATA Phase 2 (signed now)

➥ Data Management Plan <u>required</u>

Electronic Phase 0

Electronic Phase 1 and Phase 2
PRISMA, Trace, coupling to xDAQ …

Electronic Phase 2 only
Ancillaries …

GGP only

GGP only + V2

2020 | 2021 | 2022 | 2023 | 2024 | 2025 | 2026 | 2027 | 2028 | 2029 | 2030

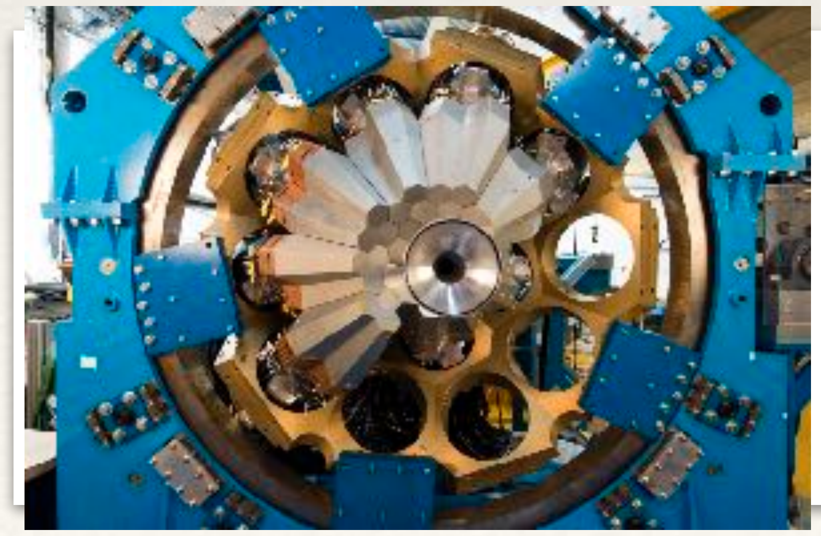
Dismounting
@ GANIL

Re-start
@ Legnaro

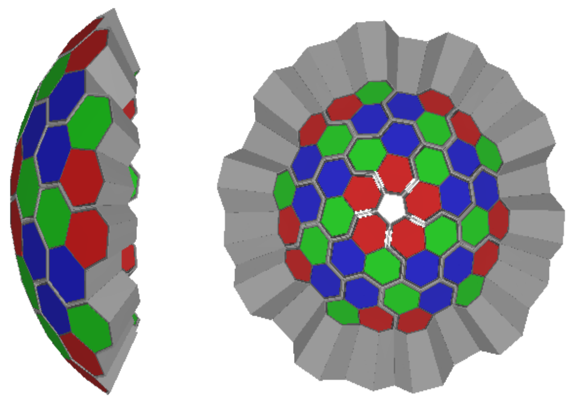
DMP ?

4 meetings

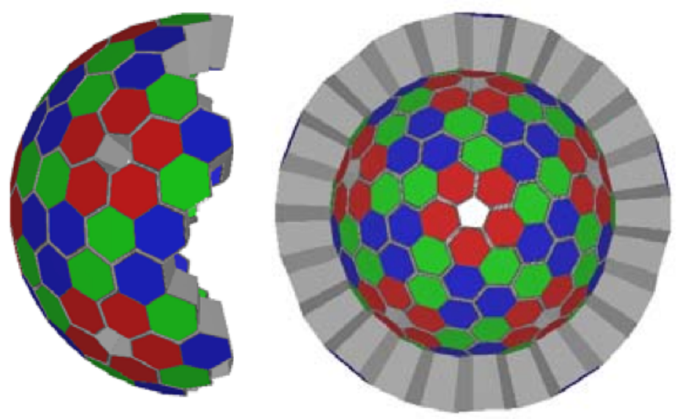
… a path to …



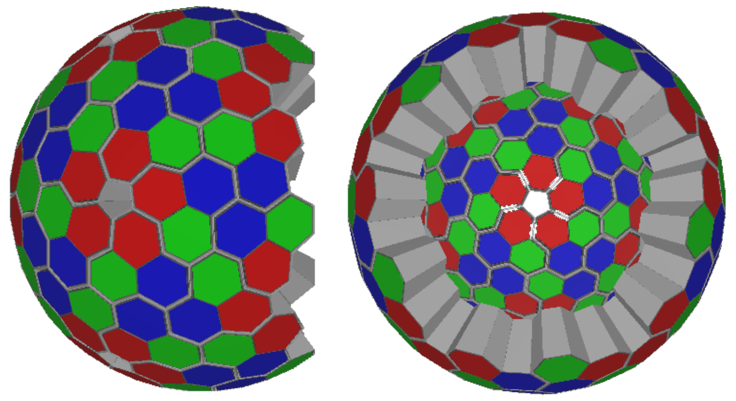
4PI/3
60 Ge
20 TC



2PI
90 Ge
30 TC



3PI
135 Ge
45 TC



# What is behind the 'DMP' door ? Open Data - Open Science Path …

Our researches are almost fully funded by **public** agencies !

*All the products from such funds have to be given back to the society* ⟶

Not only Papers !

Open Source
Open Methods
Open Peer Review
Open Data
Open Education
Open Access

4 meetings

Results of our discussions
2 documents on ATRIUM
Current Practices
What to be changed ?

Step# 0 : Where we are

(Personal wiew)

DMP

Step#1 : DMP

Open sources

Open methods

Building a level scheme
in his internet browser

citizen

Step#n: Open Science

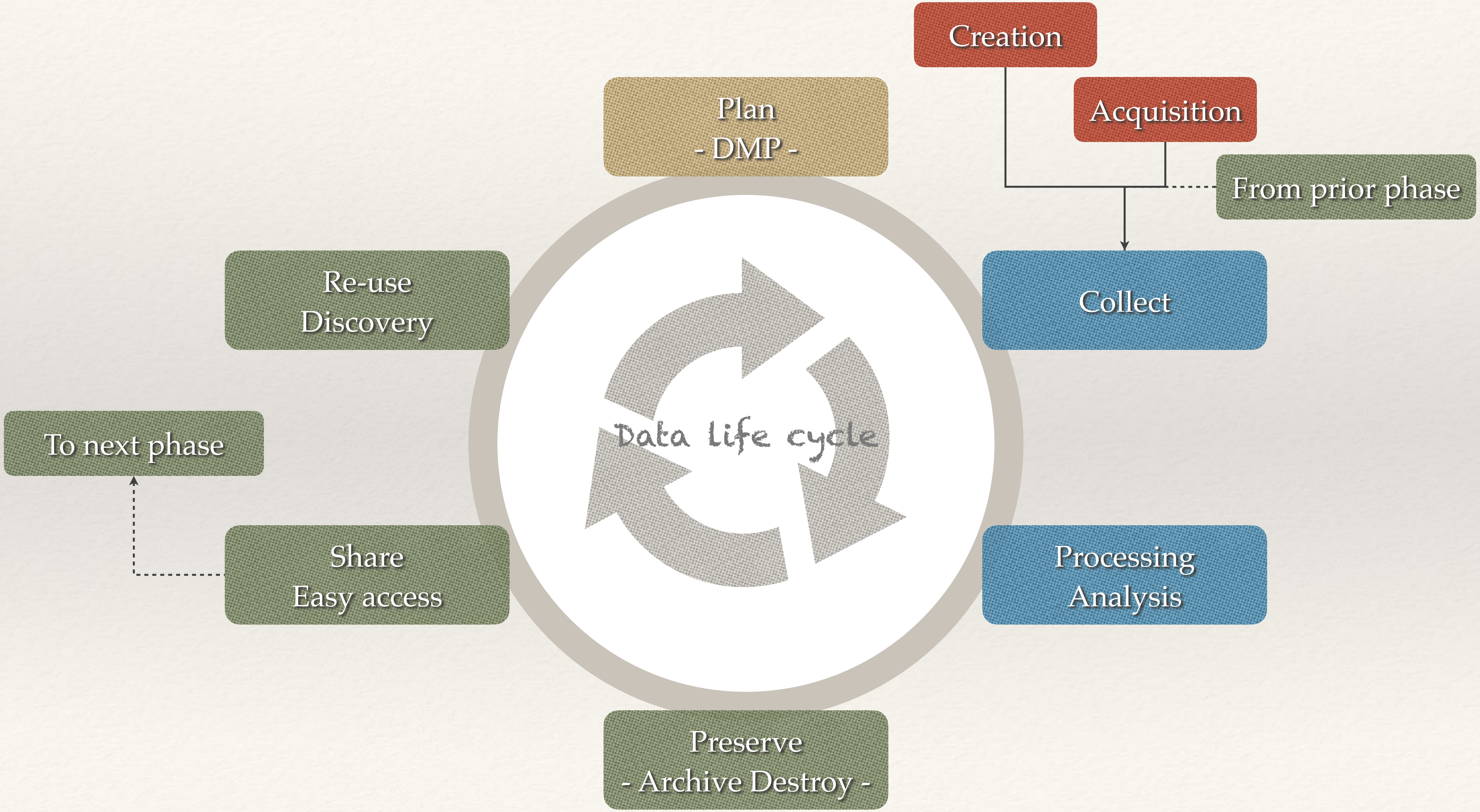
Step#i : Open Data

« data scientists »
A colleague playing with « my » data

**AGATA Collaboration Council (decisions) actor along the path**

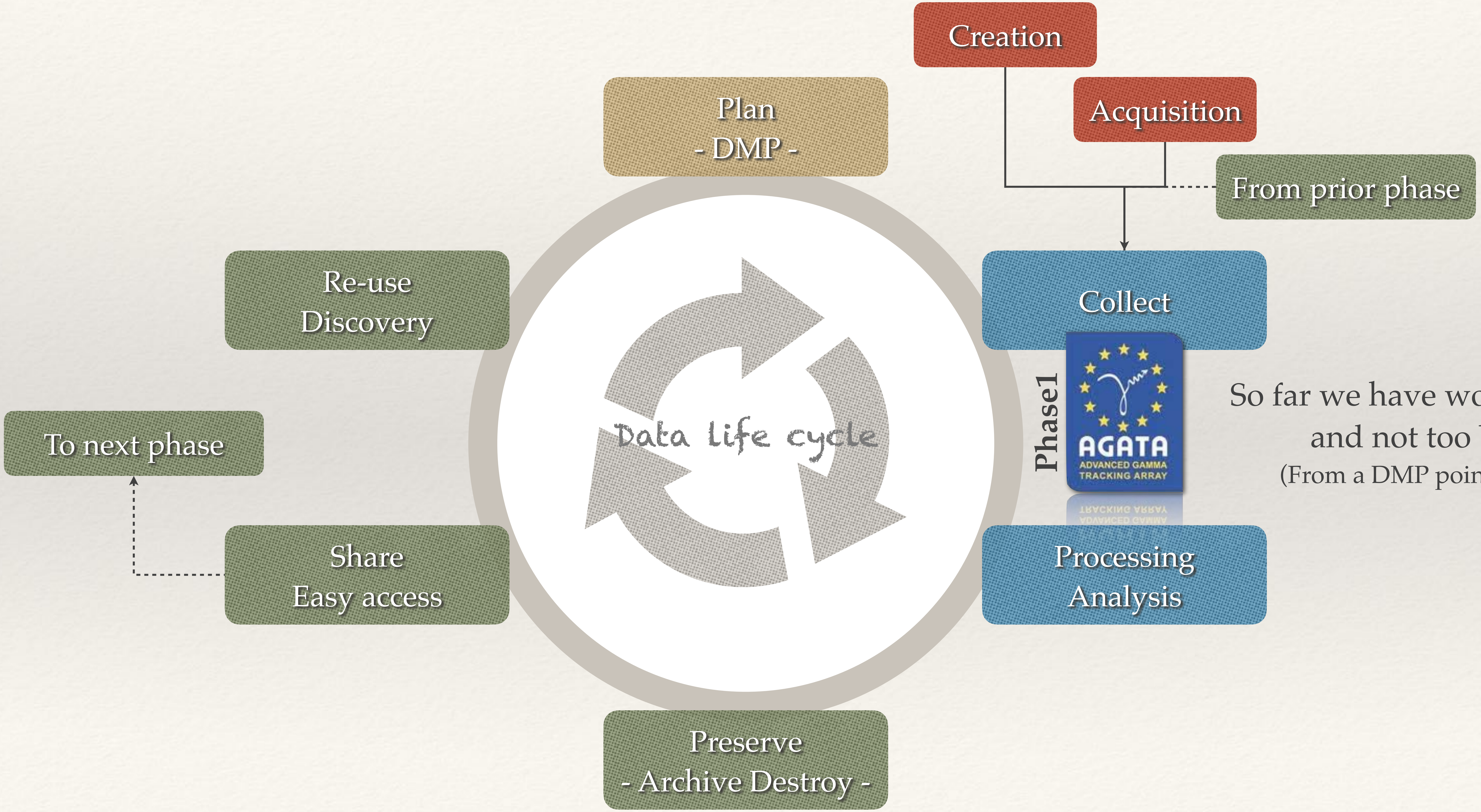
# Step #1 : what is a DMP ?

It means we should establish a plan to deal with the cycle of life of the data produced by AGATA …



# Step #1 : what is a DMP ?

It means we should establish a plan to deal with the cycle of life of the data produced by AGATA …



Creation

Acquisition

From prior phase

Plan
- DMP -

Re-use
Discovery

Collect

Phase1

AGATA
ADVANCED GAMMA
TRACKING ARRAY

Data life cycle

To next phase

Share
Easy access

Processing
Analysis

Preserve
- Archive Destroy -

So far we have worked here !
and not too badly
(From a DMP point of view)

# Step #1 : what is a DMP ?

It means we should establish a plan to deal with the cycle of life of the data produced by AGATA …

New to be added for the Phase 2 [?]

Plan
DMP -

Creation

Acquisition

From prior phase

Collect

Phase1

Data Life cycle

Re-use
Discovery

To next phase

Share
Easy access

Preserve
- Archive Destroy -

Processing
Analysis

So far we have worked here !
and not too badly
(From a DMP point of view)

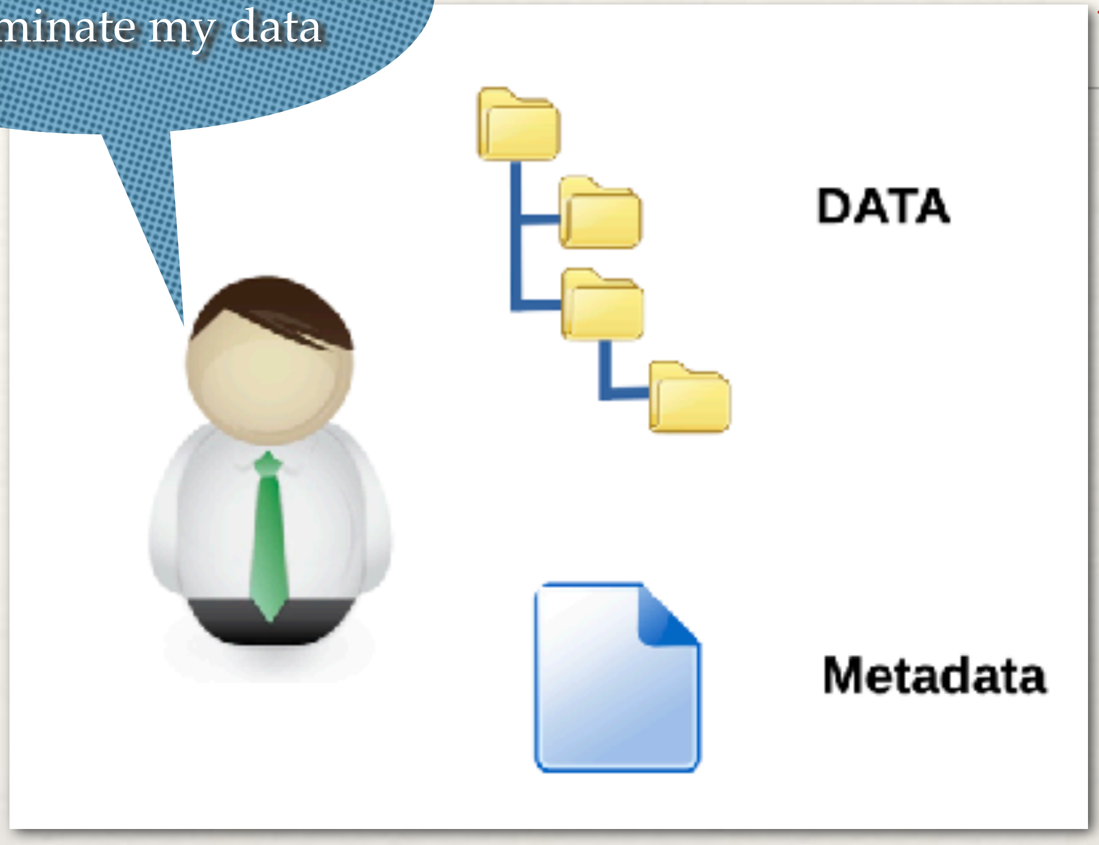
To be enhanced for the Phase 2 [?]

AGATA
ADVANCED GAMMA TRACKING ARRAY

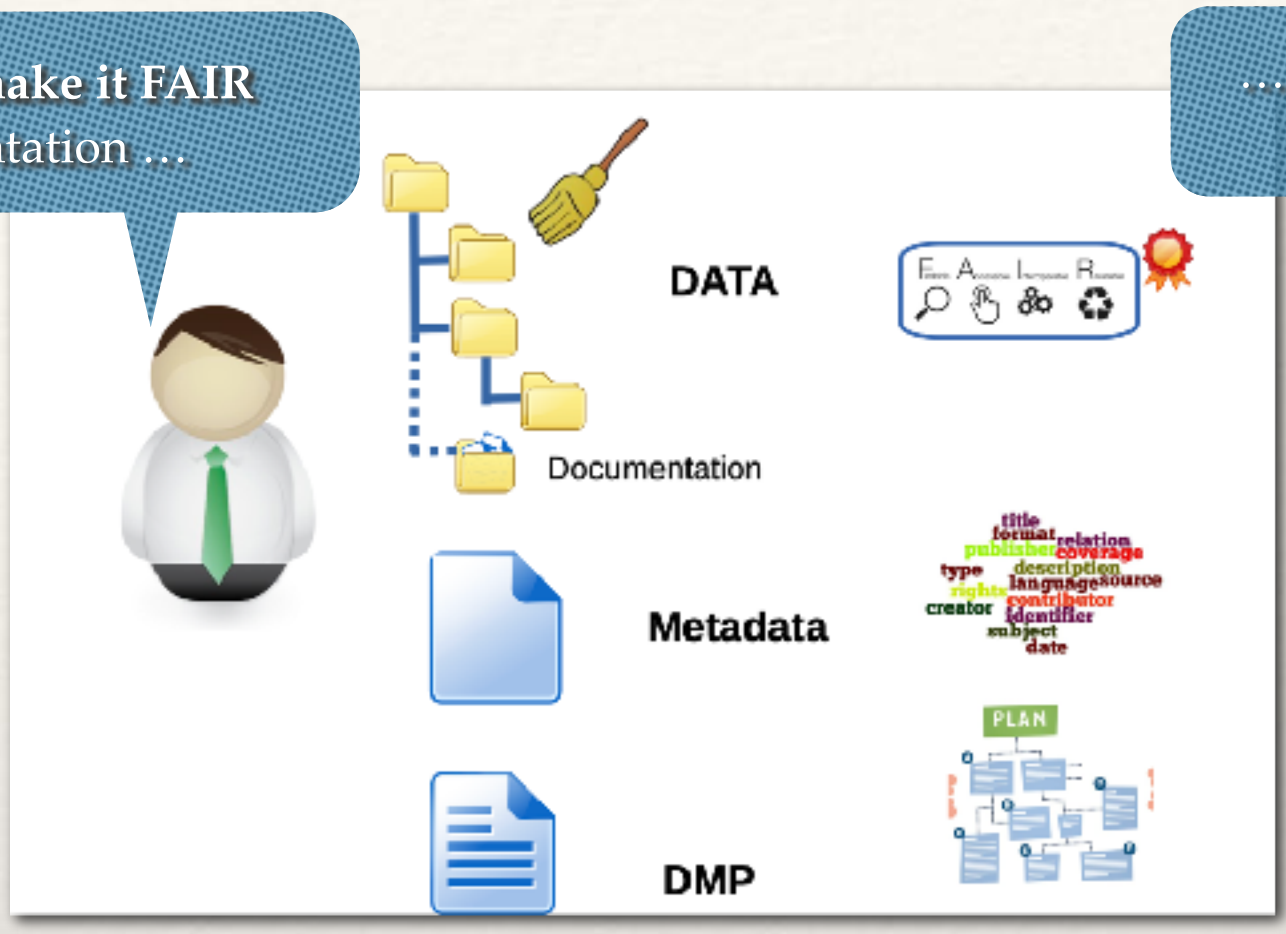
# Step #1 : what should we do ?

How to fulfil the cycle ?

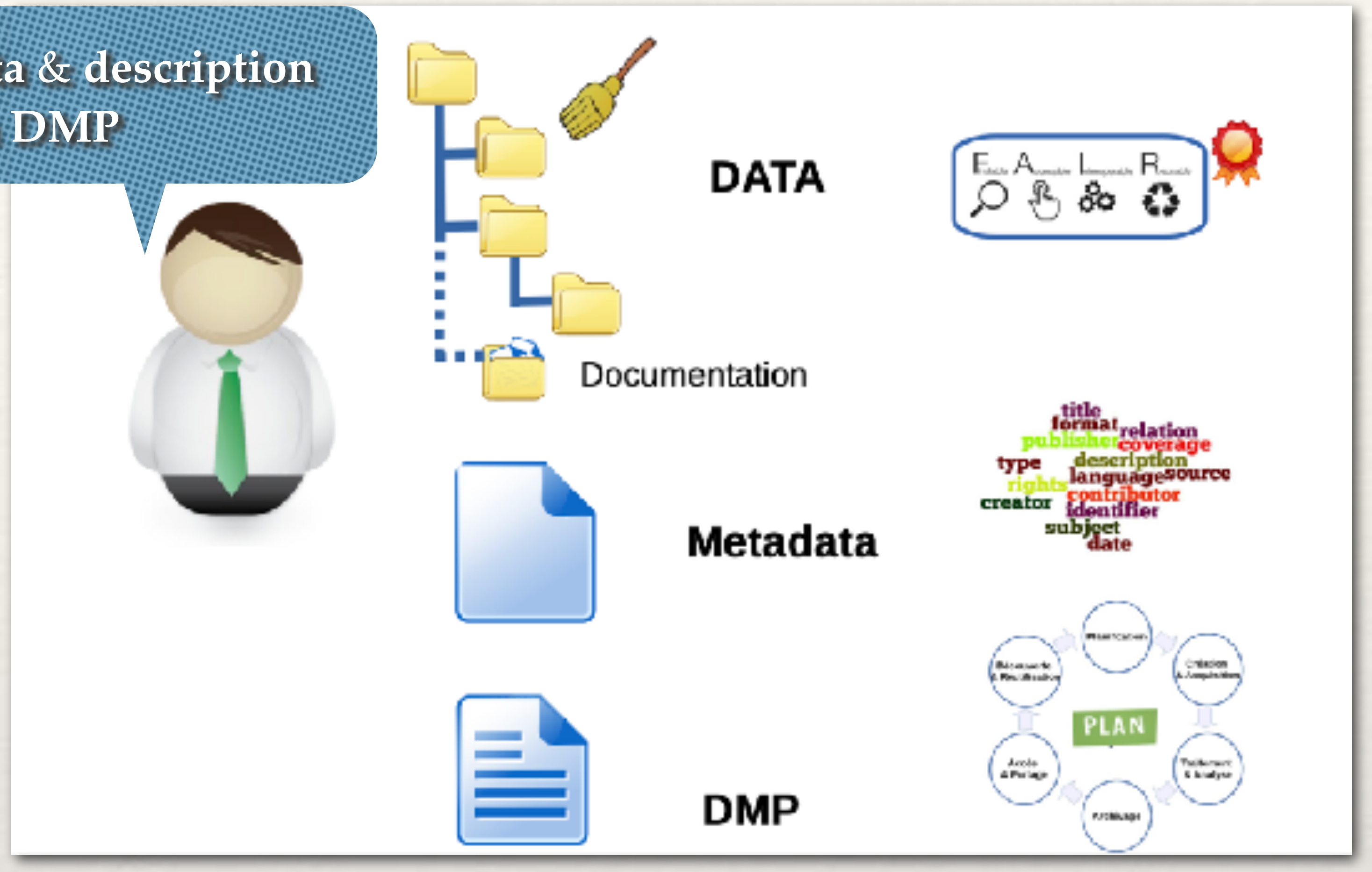
I want to archive and disseminate my data



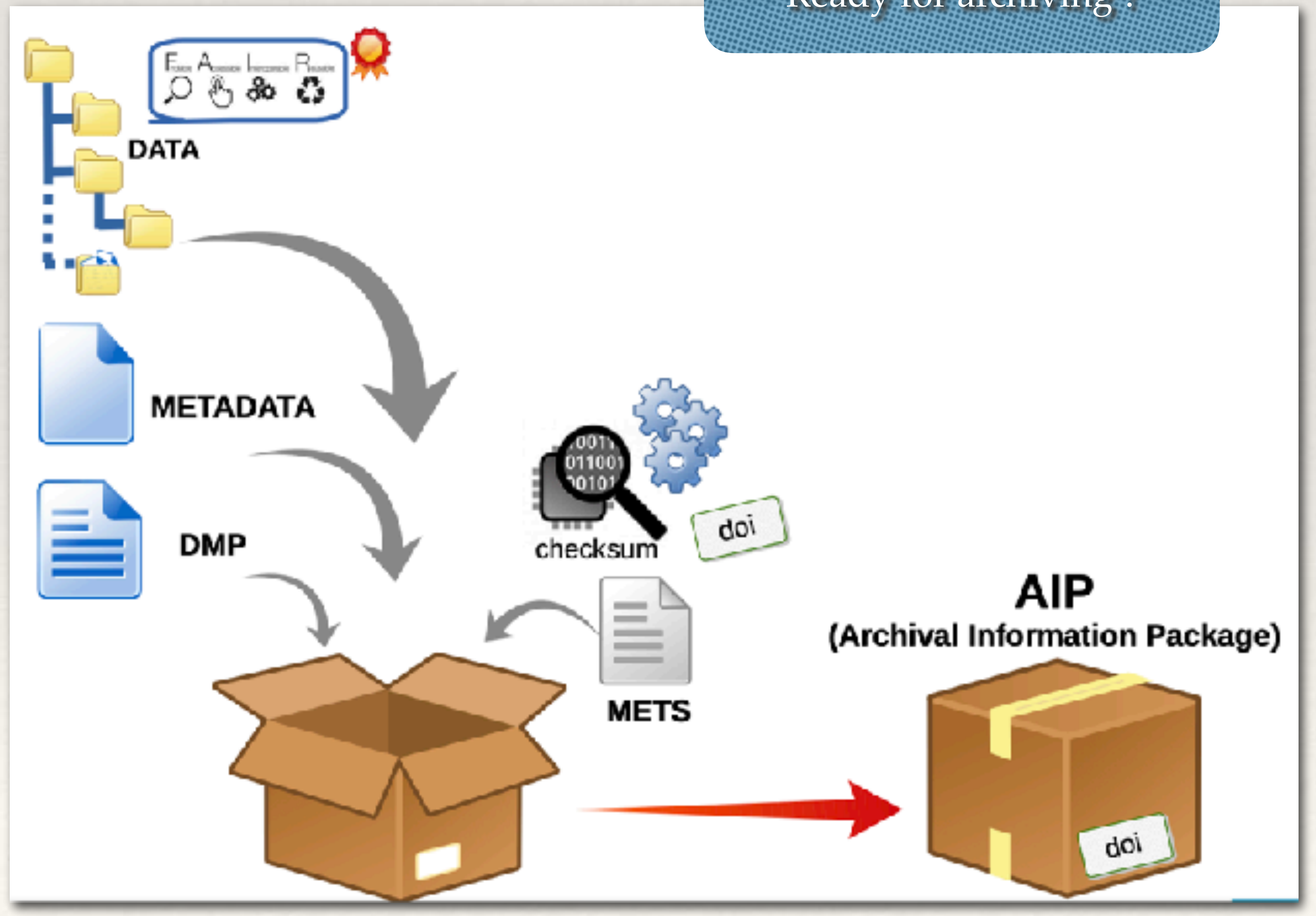
Clean the data, **make it FAIR**
Add documentation …



… **Rich meta-data & description**
Write a **DMP**

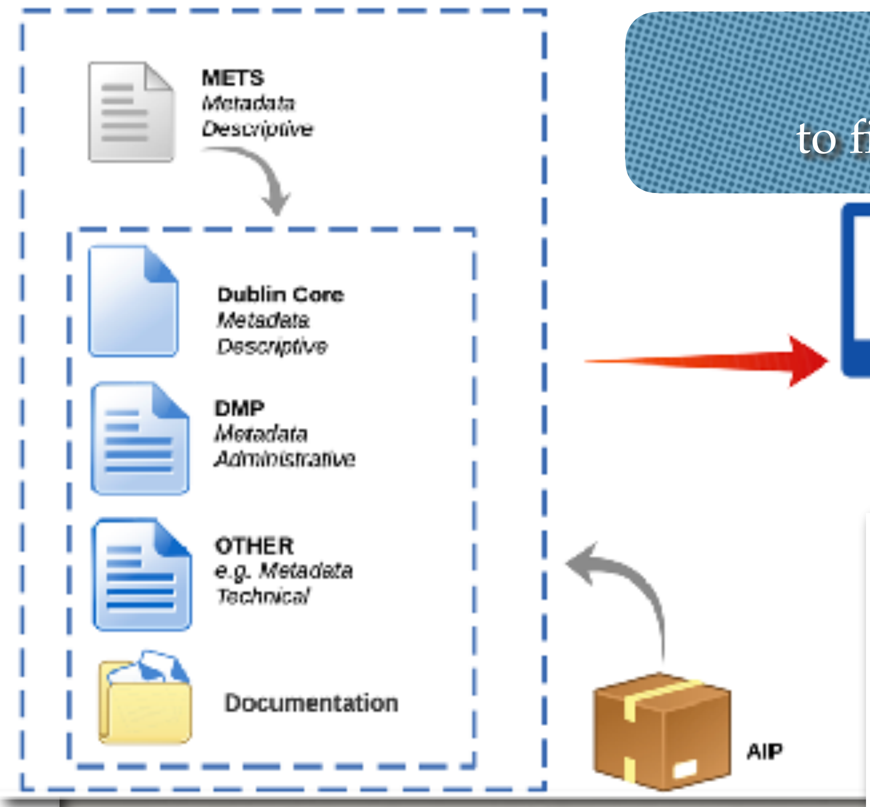


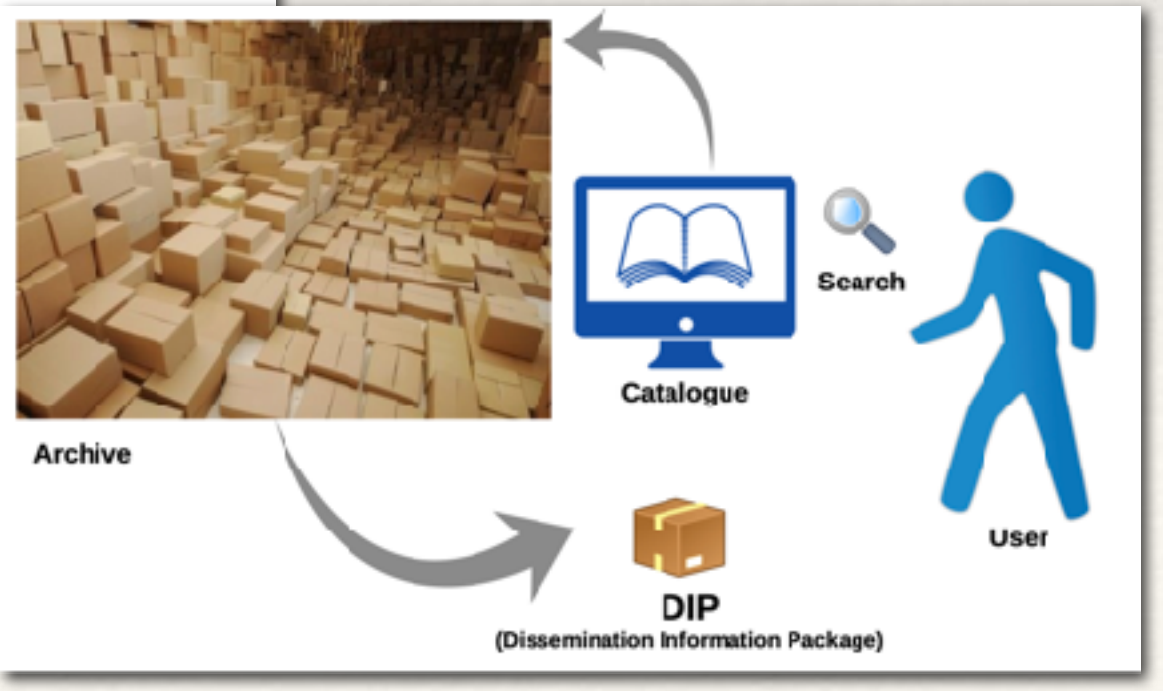
Ready for archiving ?



We don't want this …

Catalogues required
to find and get back an archive





Cartoon from this presentation

# What AGATA produced as Data ?

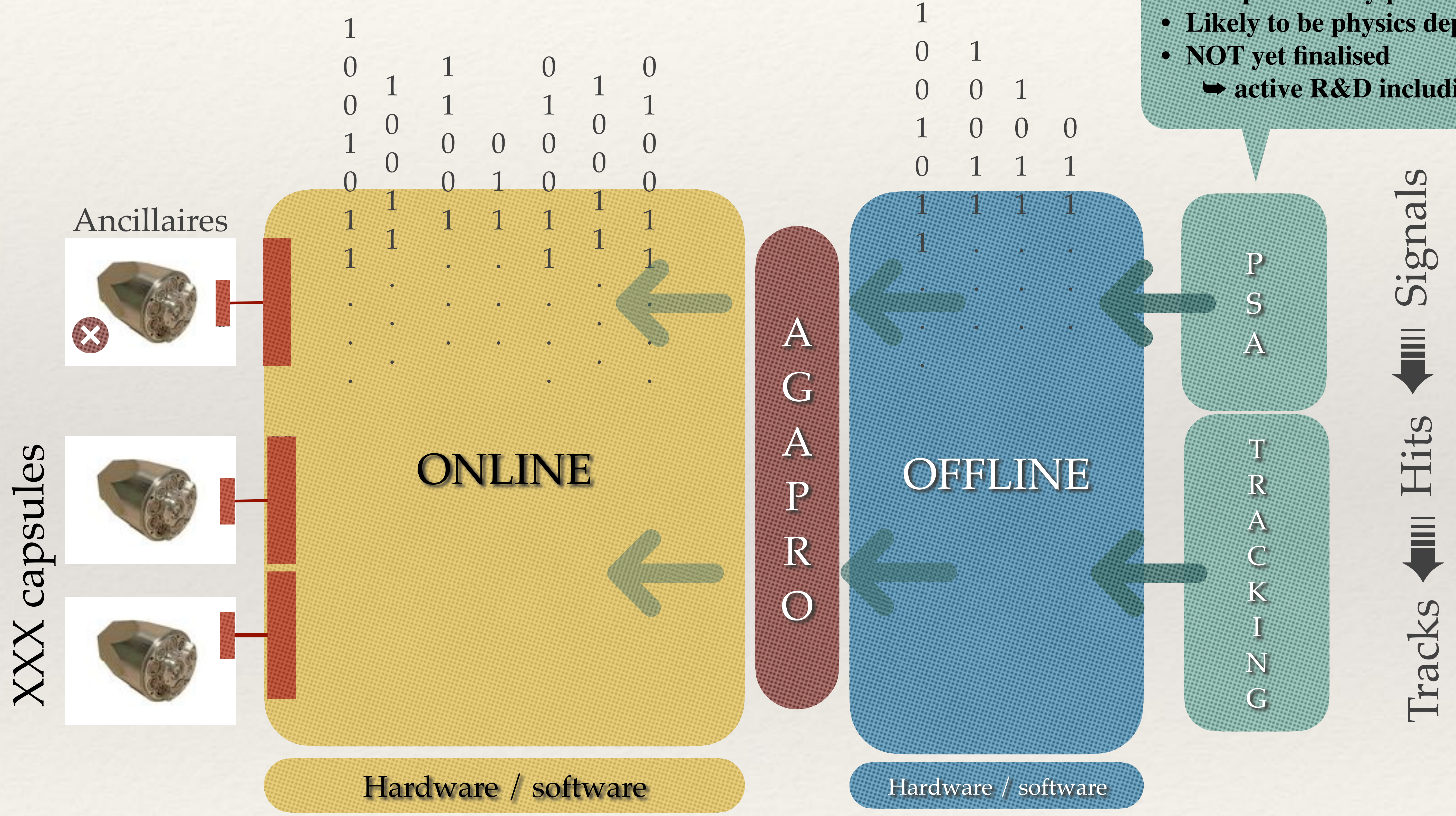
Schematic view of the Data Production

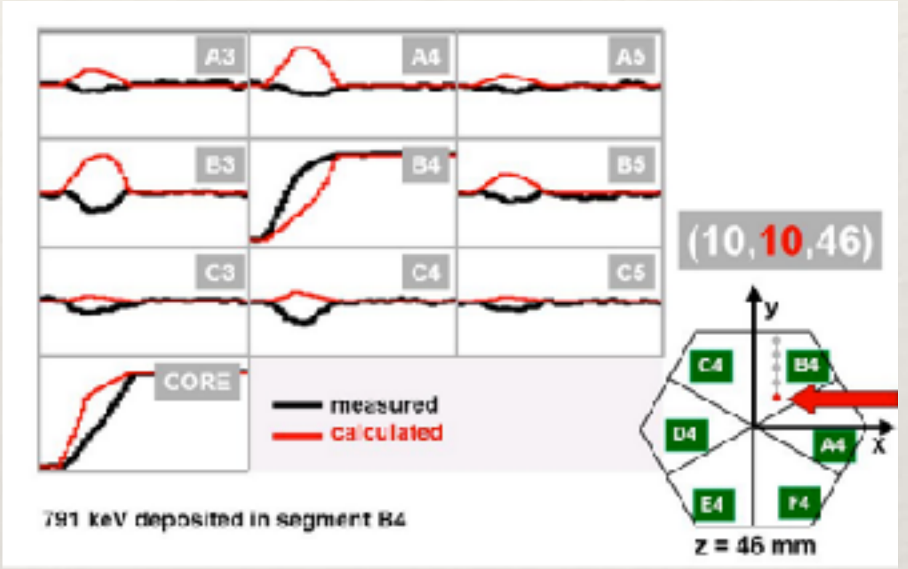
AGATA shortsighted without the 2 key algorithms

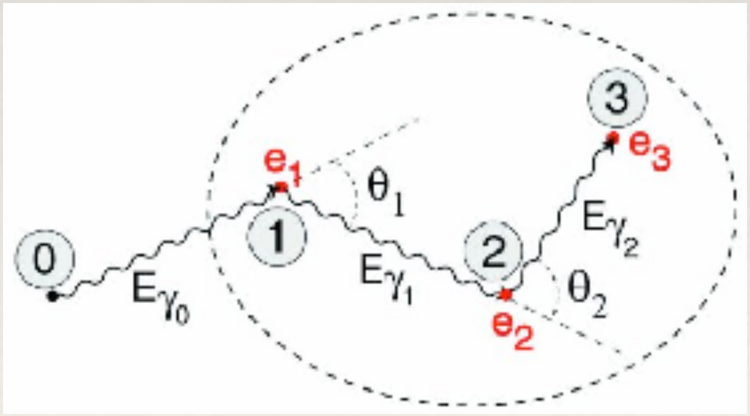
They are :

- Complex - many parameters, resources demanding, …
- Likely to be physics dependant
- NOT yet finalised

➥ active R&D including Machine Learning approaches







# What AGATA produced as Data ?

Schematic view of the Data Production

Online Data stored (for distribution) on the grid
+ part of the meta-data

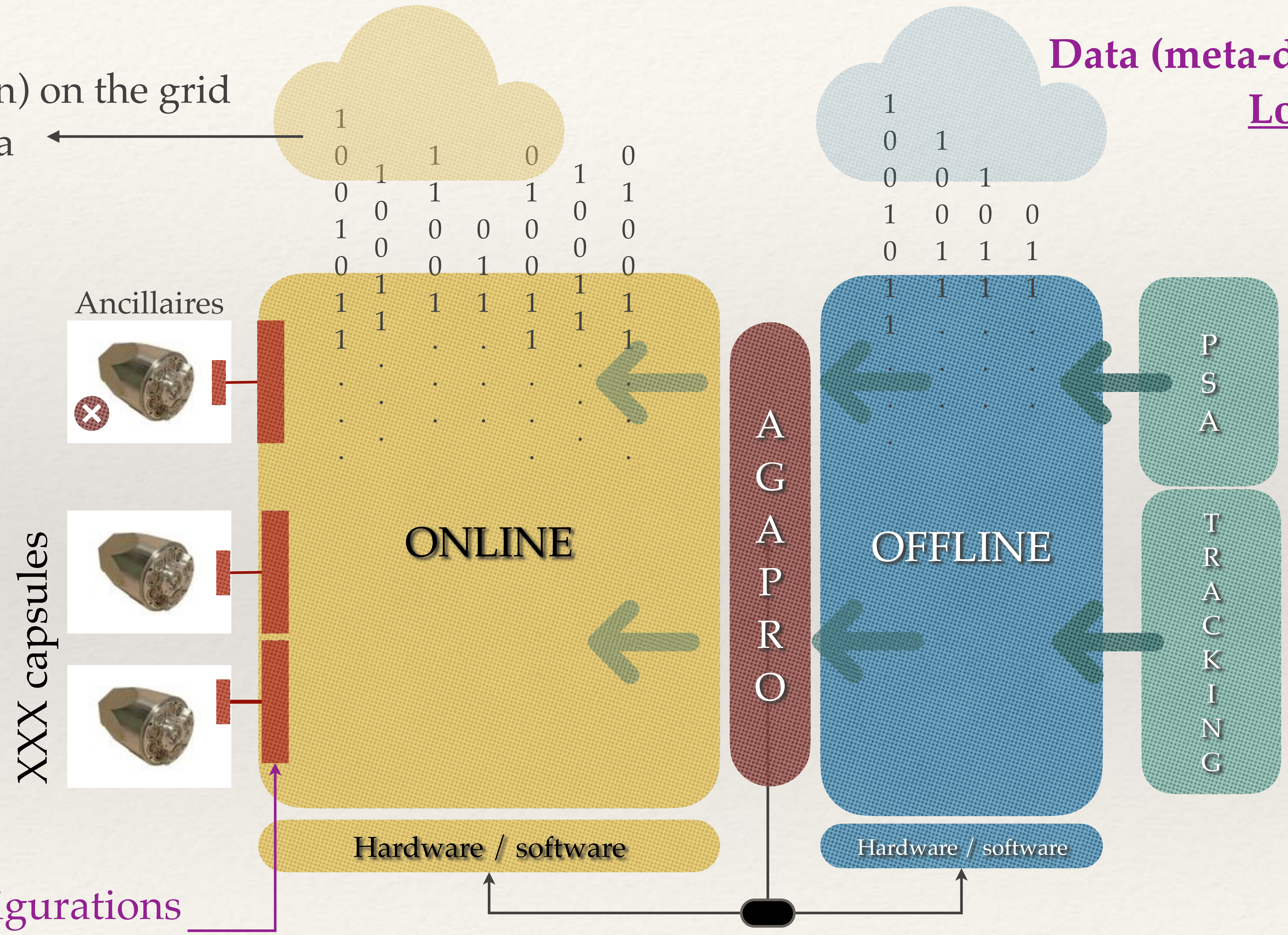
**Data (meta-data) produced offline (@home)**
**Lost or almost lost … !**

Good, with have :
- Backups, access through VO,
- A policy to use data

…

but :
- No catalog
- Poor meta-data

…



**Are**
**the simulated data set,**
**the produced analysis,**
**the software,**
**the meta data**
**…**
**saved ? … NO !!**

meta-data electronic + ? Configurations
Lost

Most of the Software in svn, git or Hg ✓
… but are the version used to produce a particular data set saved ?

# Phase 1 : our workflow, more details

Producer
Intermediary
Consumer
AD ≡ AGATA Data
AD_Level_0
Traces
AD_Level_1
Hits
Ancillary
AD_Level_0
Ancillary Data
PrePSA
PSA
PostPSA
PCI
PCI
Builder
Merger
Tracking
AD_Level_2
Correlated Hits
AD_Level_3
Correlated Hits
+
Ancillary
AD_Level_4
Tracked Gamma
+ Ancillary
+ Correlated Hits
Local Level Processing
Global Level Processing

Production online @ different levels
Level 0 requires running again an experiment !
In principle from level 0
one can reproduce offline the other levels
ONLY IF we do have all the meta-data required

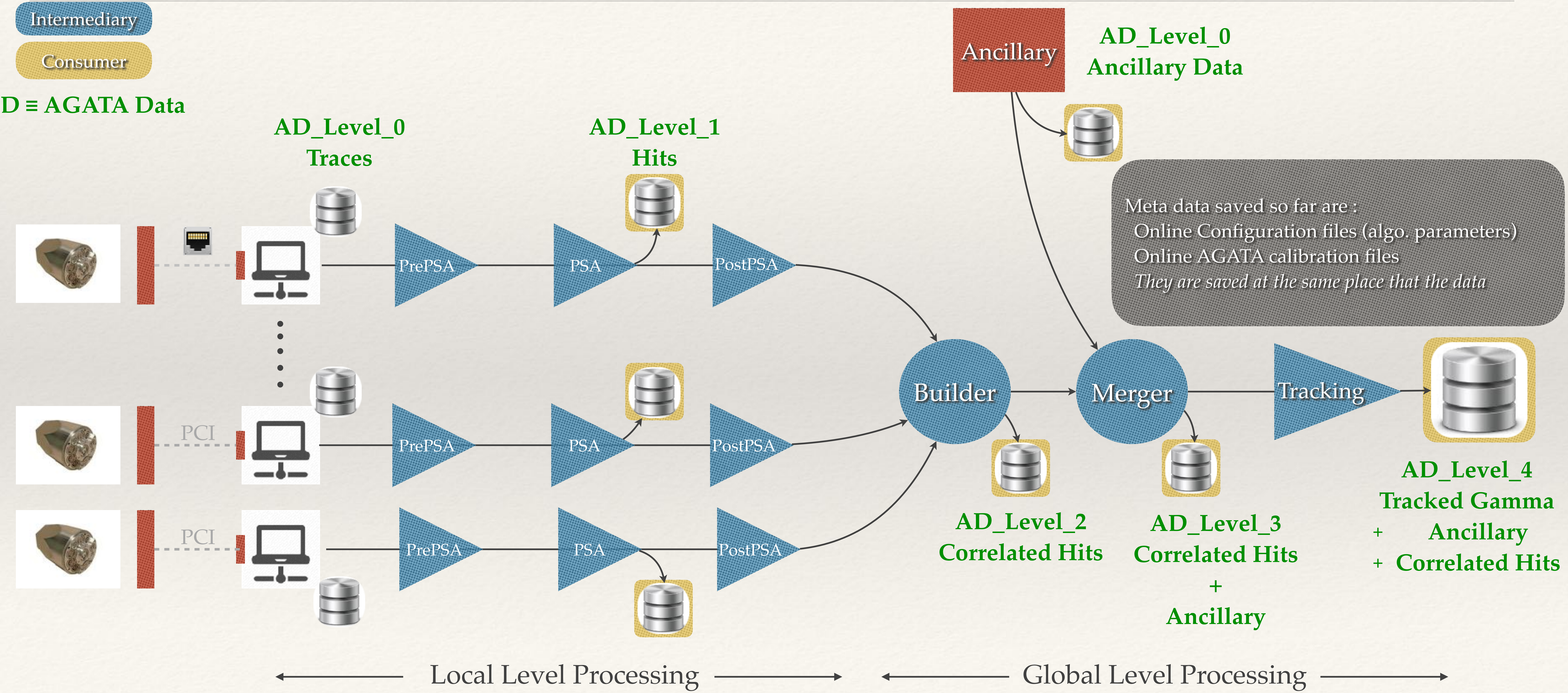
# Phase 1 : our workflow, more details

Producer

Intermediary

Consumer

AD ≡ AGATA Data



Meta data saved so far are :

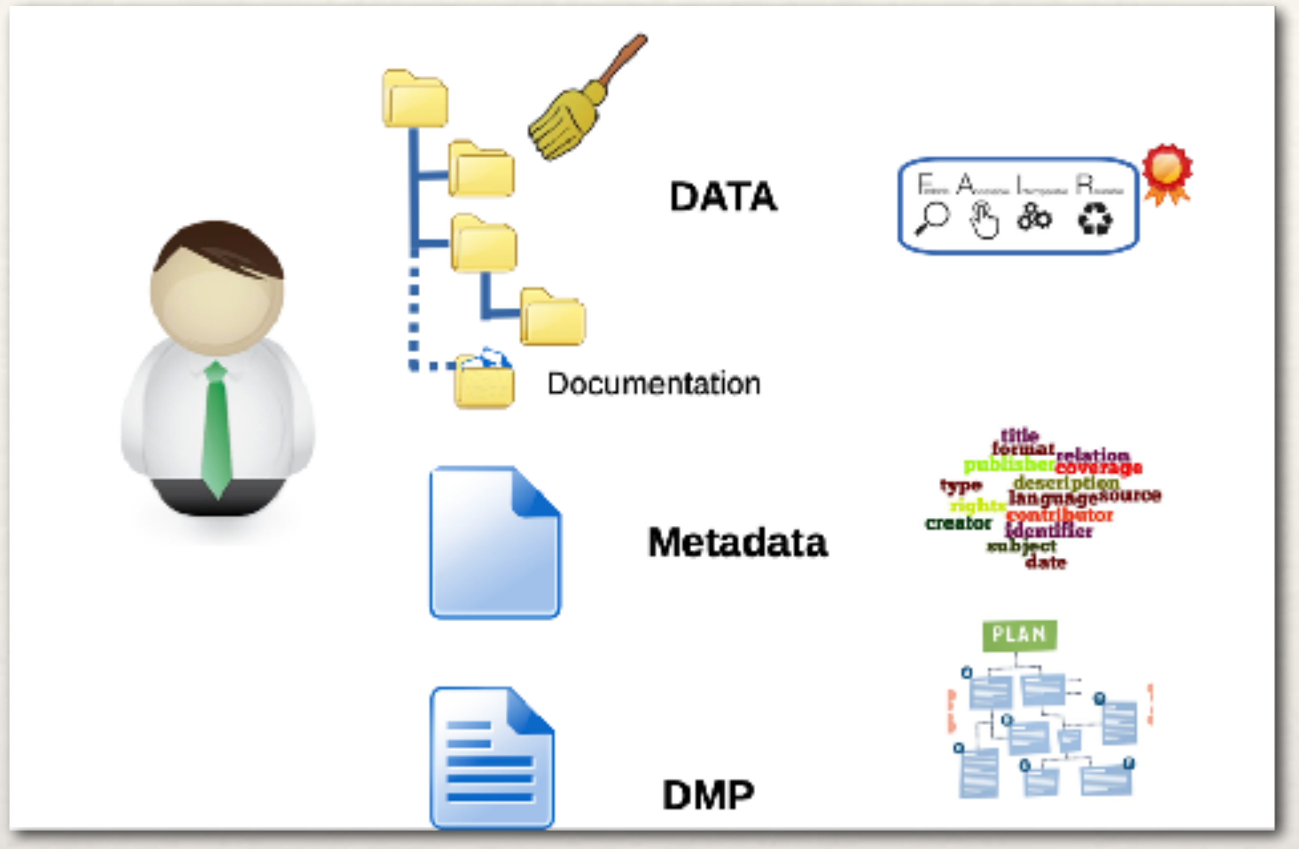
Online Configuration files (algo. parameters)

Online AGATA calibration files

*They are saved at the same place that the data*

# First step toward a DMP for the Phase 2

## FAIRification Process !!!



Cleaning, documentation … ok … but what means FAIR ???

FAIR means **F**indability, **A**ccessibility, **I**nteroperability, **R**eusability

- ➥ FAIRification process, make sure the data (+meta) produced are FAIR
  - ➩ **likely to have an impact on the way we produce, store etc … our data !**
- ➥ There are guidelines for that (see for instance https://www.go-fair.org/fair-principles/)
- ➥ let's have a look at some recommandations to be FAIR

Phase1

### Findable

F1. Data (and meta-data) are assigned a globally unique and persistent identifier (PID)*
F2. **Data are described with rich meta data**
F3. Meta data clearly and explicitly include the identifier of the data they described
F4. (Meta)data are registered or indexed in a searchable resource

* an example of PID is DOI. PID ≡ web page stored in a repository (See for instance zenodo)

F1. Obviously not the case
➥ *it might be good to start with at least a standard name for AGATA experiments*

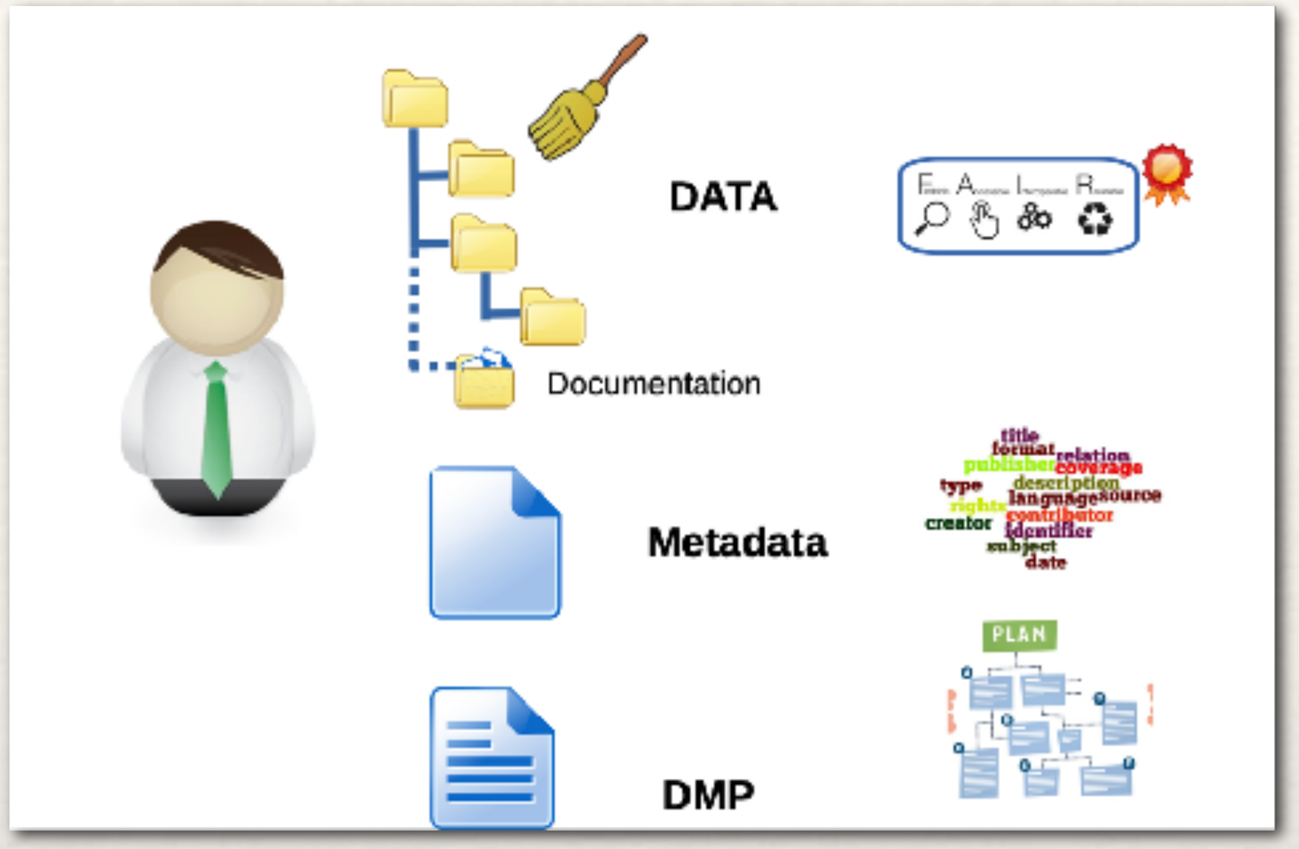
F2. We have only a minimal amount of meta data **and** only for online data

F3. Our meta data are stored inside the data …
➥ *metadata and data should be separated, see also A2*

F4 . Obviously not the case (see again zenodo)
➥ *searchable by humans and computers !*
➥ *Ex of a search: try and find all the data set produced at GANIL with NEDA ?*

# First step toward a DMP for the Phase 2

FAIRification of the data (see also the document called ADP-Part2, to be sent to the ACC)



Cleaning, documentation … ok … but what means FAIR ???

FAIR means **F**indability, **A**ccessibility, **I**nteroperability, **R**eusable

- ➥ FAIRification process, make sure the data (+meta) produced are FAIR
  - ➩ **likely to have an impact on the way we produce, store etc … our data !**
- ➥ There are guidelines for that (see for instance https://www.go-fair.org/fair-principles/)
- ➥ let's have a look at some recommandations to be FAIR

## Accessible

A1. (Meta)data are retrievable by their identifier using a standardised communication protocol
 A1.1.The protocol is open, free and universally implementable
 A1.2.The protocol, where necessary, allows for an authentification & authorisation procedure
A2. **Metadata are accessible, even when the data are no longer available**

Phase1

A1. Grid access
➥ *difficult to retrieve a particular data set without browsing all*

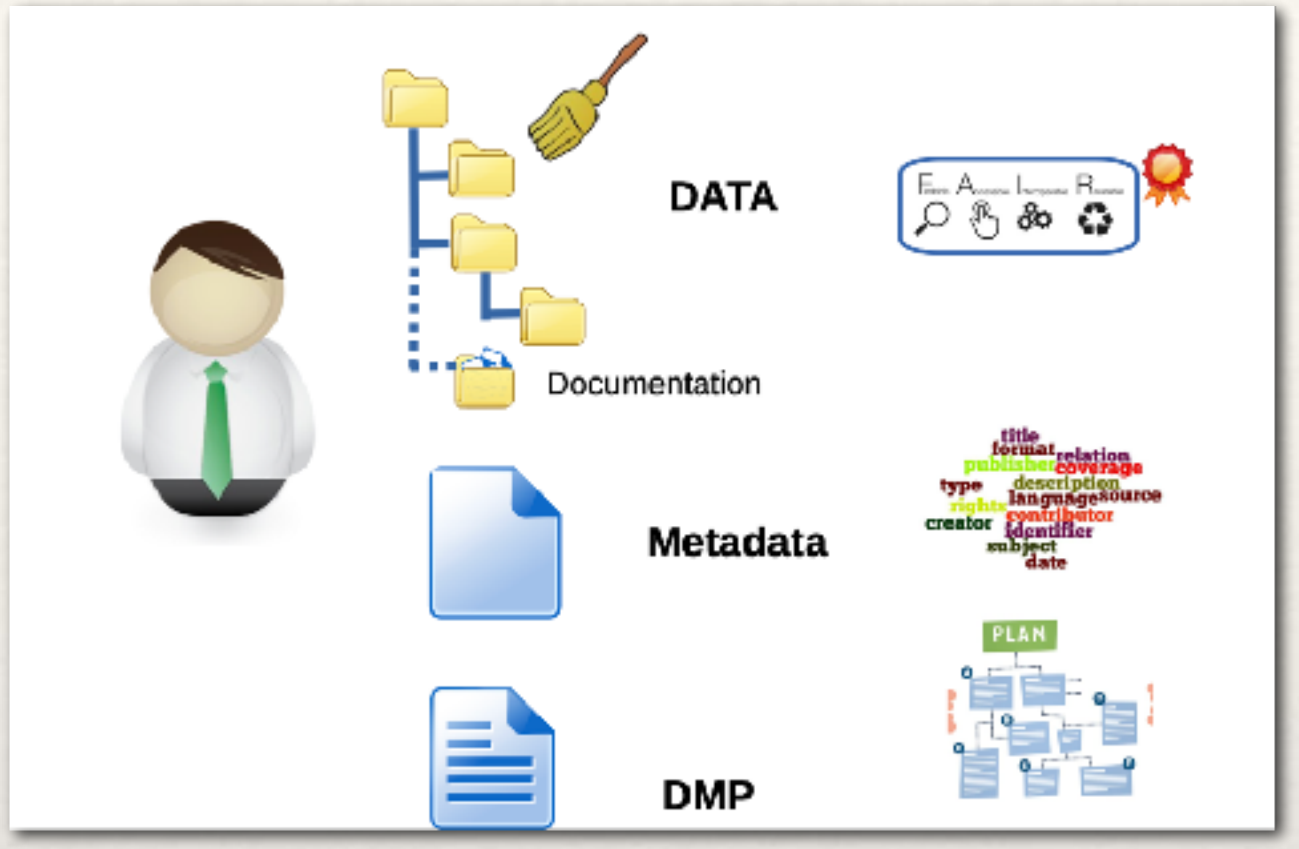
A1.1 Grid access
➥ *Not completely universal, heavy for the collaboration, time to simplify if possible*

A1.2 AGATA Virtual Organisation
➥ *is this enough ?*

A2. Obviously not the case
➥ *AGAIN meta data should be separated from data*

# First step toward a DMP for the Phase 2

FAIRification of the data (see also the document called ADP-Part2, to be sent to the ACC)



Cleaning, documentation … ok … but what means FAIR ???

FAIR means **F**indability, **A**ccessibility, **I**nteroperability, **R**eusable

➥ FAIRification process, make sure the data (+meta) produced are FAIR

⇨ **likely to have an impact on the way we produce, store etc … our data !**

➥ There are guidelines for that (see for instance https://www.go-fair.org/fair-principles/)

➥ let's have a look at some recommandations to be FAIR

## Interoperable

I1.(meta)data use a normal, accessible, shared and broadly applicable language for knowledge representation

I2. **(meta)data use vocabularies that follow FAIR principles**

I3. Meta-data qualified references to other (meta)data

This is for integration of AGATA data with other data …

**Almost nothing done so far to help in that path …**

## Re Usable

R1. (Meta)data are **richly described with a plurality of accurate and relevante attributes**

R1.1. (meta)data are released with a **clear and accessible usage licence**

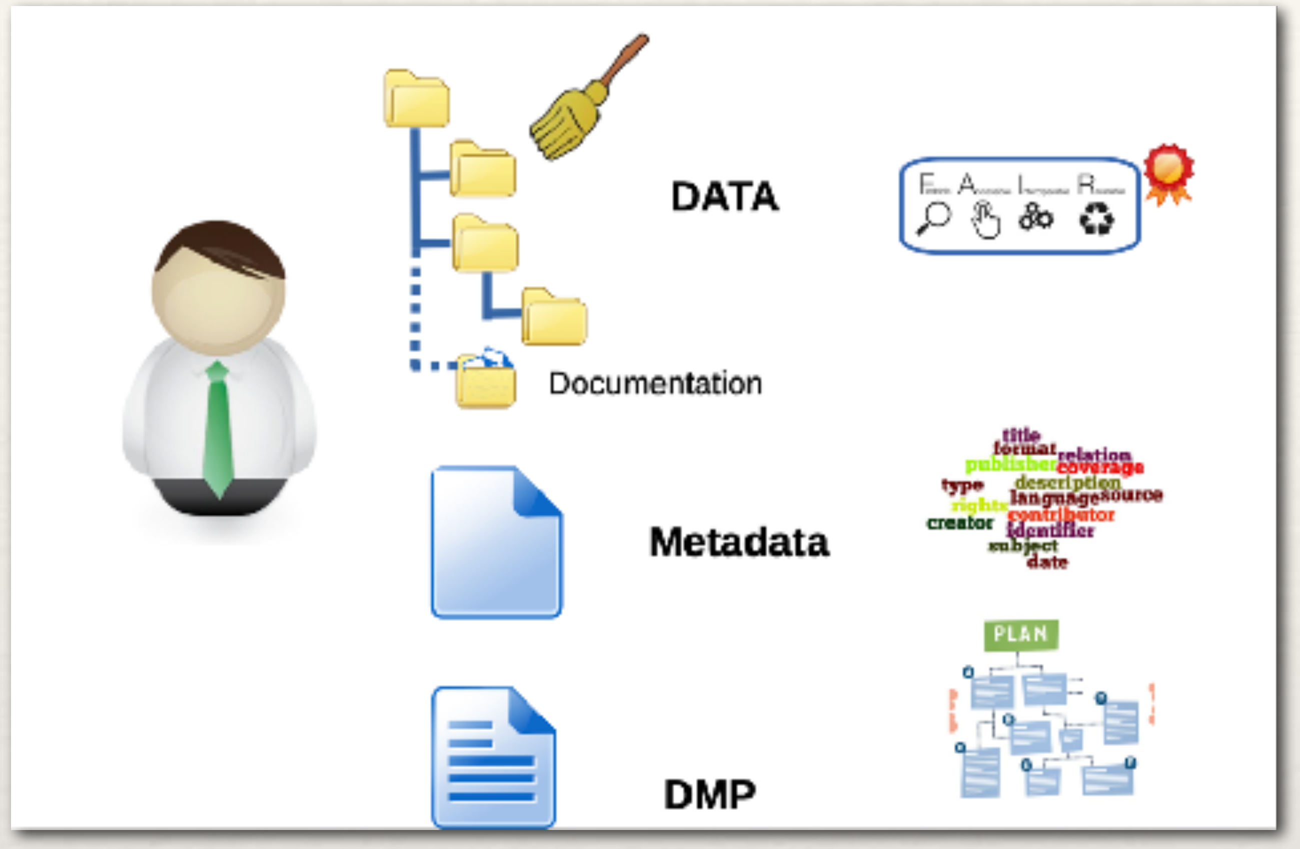
R1.2. (meta)data are **associated with detailed provenance**

R1.3. (meta)data meet **domain-relevant community standards**

This is 'others' to play with AGATA data …

**Almost nothing done so far to help in that path …**

# We do have a first DMP written …



Many guidelines to really write a DMP
Here from the French ANR agency
  All are quite similar
  Moving from one to another one should be easy

1. **Data description and collection, re-use of existing data**

A. How will new data be collected or produced and/or how will existing data be re-used?

B. What data (for example the kinds, formats, and volumes) will be collected or produced?

2. **Documentation and data quality**

A. What metadata and documentation (for example the methodology of data collection and way of organizing data) will accompany data?

B. What data quality control measures will be used?

3. **Storage and backup during research process**

A. How will data and metadata be stored and backed up during the research process?

B. How will data security and protection of sensitive data be taken care of during the research?

4. **Legal and ethical requirements, codes of conduct**

A. if personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

B. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

C. How will possible ethical issues be taken into account, and codes of conduct followed?

5. **Data sharing and long-term preservation**

A. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

B. How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?

C. What methods or software tools will be needed to access and use the data?

D. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

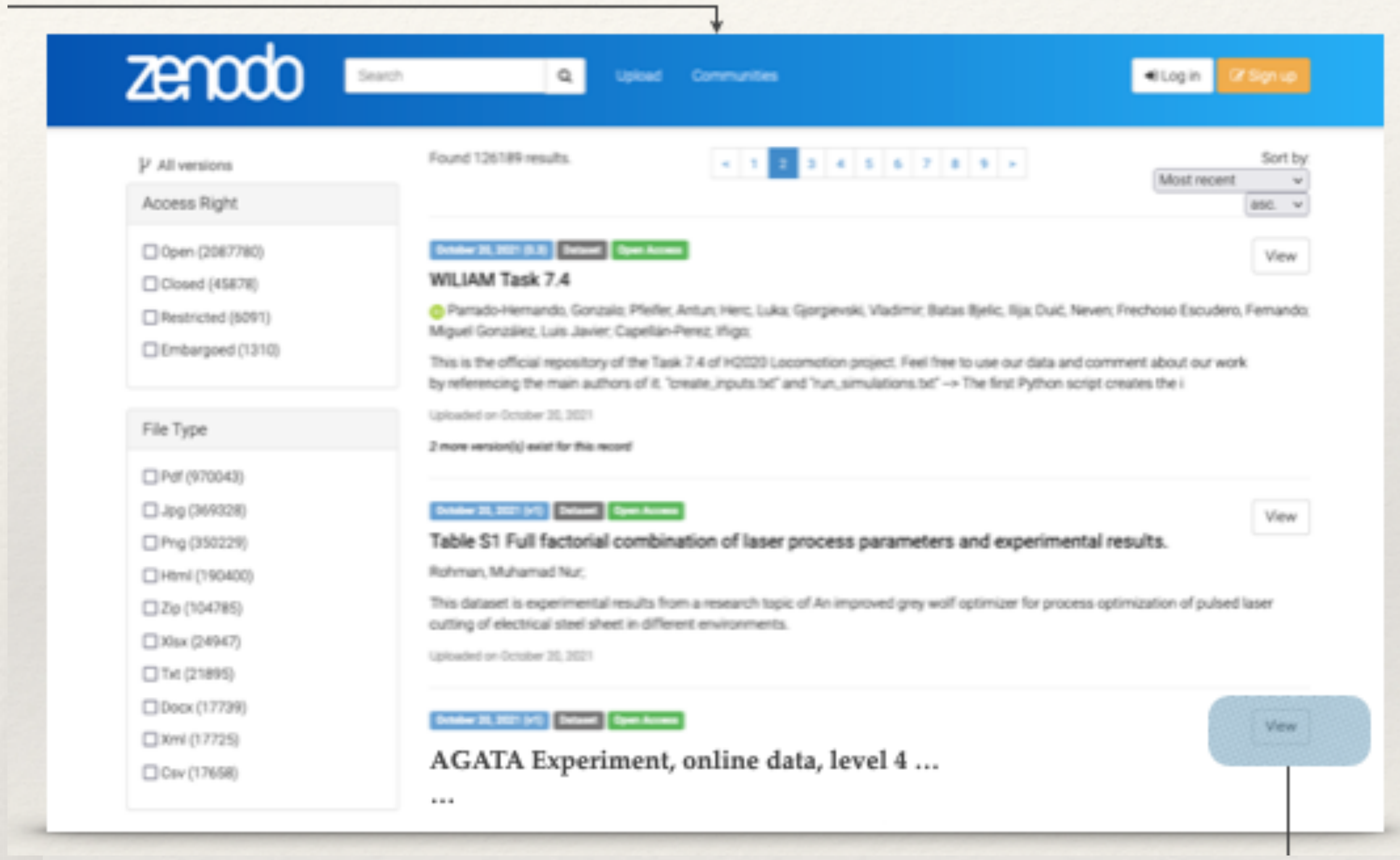
6. **Data management responsibilities and resources**

A. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

B. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

AGATA
Data Policy
for
Publications

# Conclusions



zenodo Search Upload Communities Log in Sign up

All versions

Access Right

- □ Open (2087780)
- □ Closed (45878)
- □ Restricted (6091)
- □ Embargoed (1310)

File Type

- □ Pdf (970043)
- □ Jpg (369328)
- □ Png (350229)
- □ Html (190400)
- □ Zip (104785)
- □ Xlsx (24947)
- □ Txt (21895)
- □ Docx (17739)
- □ Xml (17725)
- □ Csv (17658)

Found 126189 results.

« 1 2 3 4 5 6 7 8 9 »

Sort by: Most recent, asc.

Dataset Open Access View

**WILIAM Task 7.4**

Parrado-Hernando, Gonzalo; Pfeifer, Antun; Herc, Luka; Gjorgjevski, Vladimir; Batas Bjelic, Ilija; Duić, Neven; Frechoso Escudero, Fernando; Miguel González, Luis Javier; Capellán-Perez, Iñigo;

This is the official repository of the Task 7.4 of H2020 Locomotion project. Feel free to use our data and comment about our work by referencing the main authors of it. "create_inputs.txt" and "run_simulations.txt" --> The first Python script creates the i

Uploaded on October 20, 2021

*2 more version(s) exist for this record*

Dataset Open Access View

**Table S1 Full factorial combination of laser process parameters and experimental results.**

Rohman, Muhamad Nur;

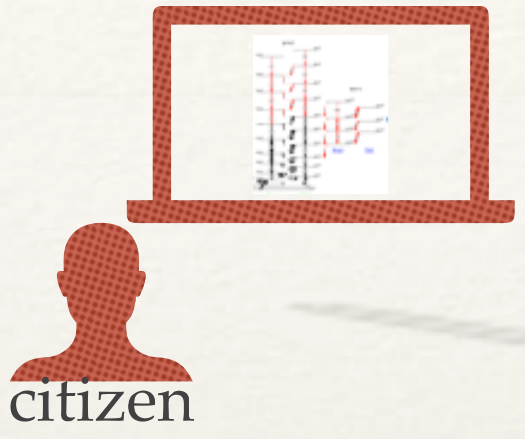
This dataset is experimental results from a research topic of An improved grey wolf optimizer for process optimization of pulsed laser cutting of electrical steel sheet in different environments.

Uploaded on October 20, 2021

Dataset Open Access View

**AGATA Experiment, online data, level 4 …**

…



Docker

First test MongoDB

Rucio* ?

Production of data bases for meta data + catalogs

Json instead of user's format

Meta-data work : move to standard file format

❷ We need to draw it using the DMP (A DMP is regularly modified)

We do not start from 0, we have identified good / bad practices

☞ progressive modification (FAIRification) of our way to work

❶ We are at the beginning of the 'open' path

* https://rucio.cern.ch/

# Thank you for listening

Questions ?

# Phase 1 : our practices so far …

+ femul on grid

R&D PSA - Tracking

Ancillaires

DMP

PC

Emulator [femul] - Offline ‘infrastucture’ - NO REAL TIME

Zone Partagée

V2

V1

V0

PSA

Tracking

+

Ancillaire

Analyses de Physique

Cluster

DAQ BOX [DCOD] - online ‘infrastructure’ - REAL TIME

Electonique AGATA

Host laboratory network

PSA / Tracking

Many Ancillaries